A Study on Prediction of Travel Time over Intervals between Adjacent Bus Stops Using Probe Data.

マンスル, アス

KYUSHU UNIVERSITY

GRADUATE SCHOOL OF INFORMATION SCIENCE
AND ELECTRICAL ENGINEERING

A Study on Prediction of Travel Time over Intervals between Adjacent Bus Stops Using Probe Data

---

Author:

Mansur As

Supervisor:

Professor Tsunenori Mine

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Engineering (Computer Science) in the

DEPARTMENT OF ADVANCED INFORMATION TECHNOLOGY

Japan

January 25, 2019

Abstract

Mansur As

A Study on Prediction of Travel Time over Intervals between Adjacent Bus Stops Using Probe Data

Prediction of travel time over travel routes is an important factor for many people who travel or commute, especially using the public bus services. Mostly, people consider travel time as well as stochastic factors in several time periods such as dwell time, traffic congestion, accidents, and so on. They usually prefer to minimize the time spent traveling from their origin to destination by choosing the best route. Intelligent Transportation Systems (ITSs) have recently been widely used in planning, evaluation and control of the reliability of the public transportation system. One of the evolutions in ITSs is largely related to bus probe data. Probe data, which are generated by vehicles, include data obtainable from navigation systems, such as time and position (longitude and latitude), i.e. data on the vehicle's running and performance history. In addition, using bus probe data enables us to monitor the state of road traffic characteristics and to measure the reliability of and variability in travel times of public bus services.

In this research, I have proposed nonlinear dynamical methods for predicting bus travel time over each interval between adjacent bus stops for seven and eight time periods in a day. The proposed methods basically utilize time series methods based on machine learning techniques: Artificial

Neural Network (ANN), Support Vector Machine (SVM) and Random Forest (RF). For this purpose, at first, I classified intervals between adjacent bus stops into two classes: stable and unstable. Then I identified two statistically significant factors: variabilities in travel time in the same time periods over days and correlations of travel time between eight time periods, which influence bus travel time over unstable intervals in the current time period. I conducted experiments to evaluate our proposed methods. I have used real bus probe data collected from November 21st to December 20th, 2013 and provided by Nishitetsu Bus Company, Fukuoka, Japan. I summarize the results obtained in this research as follows:

- Chapter 1 describes background information concerning Intelligent Transportation System (ITSs), including Bus Probe Data as a robust data source for predicting bus travel time, problem formulation, uniqueness of this research, the framework and the objectives of this research.

- Chapter 2 presents a literature review regarding prediction of bus travel time. A variety of models and algorithms have been developed to predict bus arrival times or bus travel times, and these can be classified into the following categories: historical average models, regression/time series models, and machine learning techniques including Artificial Neural Network (ANN) Support Vector Machine (SVM) and Random Forest (RF).

- Chapter 3 explains the data used in this research, describes the preliminary data preparation for the analyses and the methods used in prediction models. In addition, I describe how real-world bus probe data can be utilized, what kind of analytic results are obtained, how the data should be calculated and what kind of challenges are involved when handling this bus probe data.

- Chapter 4 discusses some of the highlighted analytic results generated from the probe data. I first calculated the bus travel time over each interval between adjacent bus stops. Then, I distinguished the intervals into stable or unstable ones. Next, I conducted statistical analysis on travel time especially over unstable intervals considering the variations in the travel time between different time periods in a day, in the same time periods over the past days and the correlation of travel time between adjacent time periods in a day. Based on the results of these analyses, I employed two types of input data: Dynamic Average Travel Time (DATT) and Historical average travel time (HATT). DATT denotes the average travel time in the time period right before the current one. Introducing DATT can be expected to adjust the error in predicting travel time in the current time period using the travel time observed in the period just before the current time period. HATT denotes the average travel time in the same time period during the past several days. It is a very important input variable of the model because HATT is more effective when travel time tends to be more consistent over days.

- Chapter 5 proposes nonlinear dynamical models for predicting bus travel times using historical information. I developed time series prediction models based on Artificial Neural Network (ANN), Support Vector Machine (SVM) and Random Forest (RF) to predict travel time over unstable intervals. Focusing on recurrent and non-recurrent variabilities in travel time over the unstable intervals, I predicted travel time for seven time periods in a day (omitting EM) and compared the experimental results of the various approaches.

- Chapter 6 proposes a different approach for predicting bus travel time considering traffic congestion. This approach uses two input variables,

DATT and HATT, selectively with a distinction between peak-hour and off-peak periods.

- Chapter 7 defines the role of these approaches in the evaluation of the performance of the prediction models. I compare the proposed model mentioned in Chapter 6 with the model mentioned in Chapter 5. In addition, to measure the significant effects of variables other than the single independent variable, I conduct an experiment comparing our models and the model proposed by another study.

- Finally, I conclude the thesis in Chapter 8. I describe the results of experiments and the findings of this research, and discuss some topics for future scientific research in the prediction bus travel time.

Acknowledgements

First of all, syukran wal hamdulillah thanks to Allah Subhana Wataalah, who gave me the strength and knowledge to accomplish this work.

I would like to express my deepest appreciation to **Professor Tsunenori Mine**, for the crucial and patient help provided in developing this study. Most of this work was developed thanks to challenging objectives that he set throughout my research. Without his knowledge, patience, and support this dissertation would not have been possible. The knowledge I gained from him will definitely influence the rest of my life. I am incredibly fortunate to have had him as an advisory under his tutelage and diligence all these years.

I would like to thank also **Professor Akira Fukuda** and **Professor Kenji Hisazumi**, as advisory committee members, for their comments and suggestions feedback were valuable in my PhD thesis.

I thank my friend **Dr. Eng. Nakamura Hiroyoki**, first year in computer science, for his guidance for preparation bus probe data at the beginning of this study.

Sincere thanks to all laboratory members especially, **Ms. Aya Miura, Mr. Kohei Yamaguchi**, and **Mr. Tsubasa Yamaguci** for their help in the laboratory during my PhD study with their loving support, understanding and encouragement.

Thanks to the Nishitetsu Bus Company Fukuoka Japan, for providing the bus probe data for this research project.

At the end, I must offer my deep gratitude to my parents, brothers and my sister for their unconditional support during my study abroad. They have always been my source of enthusiasm. Special thanks to my wife and my son who enriched me with love, strength, and motivation.

Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
AVI	Automatic Vehicle Identification
APTIS	Automatic Passenger Ticket Issue
GPS	Global Positioning System
DATT	Dynamic Average Travel Time
HATT	Historical Average Travel Time
ITS	Intelligent Transportation System
E	Early Morning
MP	Morning Peak
LM	Late Morning
MD	MidDay
EA	Early Afternoon
AP	Afternoon Peak
E	Evening
LN	Late Night
SVM	Support Vector Machine
SVR	Support Vector Machine Regression
ANN	Artificial Neural Network
RF	Random Forest
NARX	Nonlinear Autoregressive Network with eXogenous Inputs
DT	Decision Tree
MSE	Mean Square Error
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Square Error
MLP	Multilayer Perceptron
NARMAX	Nonlinear Autoregressive Moving Average with eXogenous
RBF	Radial Basis Function
ARIMA	Autoregressive Integrated Moving Average

List of Symbols

Tt	Travel Time
tp	Time Period
ln	Natural Logarithm
d	day on weekday
i	Interval between Two Bus Stops
N	Number of Intervals
$StDev$	Standard Deviation
μ	Average of Standard Deviations
$VarStDev$	Variance of Standard Deviations
σ	Standard Deviation of the Standard Deviation
xt	Time Series
n_{sv}	Number of Support Vector
C	Constant
u	Random Vector
K	Element of Regression Function
w	Weights of the Network
x_i	Input Vector
y_i	Scalar Output

Chapter 1

Introduction

This chapter introduces the background of this thesis. It discusses background, problem formulation, the uniqueness of this research, research objectives, and the framework of the research.

1.1 Background

Prediction of travel time on the route is an important factor for many people who travel or commute, especially using the public bus services. Mostly, passengers consider travel time as well as stochastic factors (variability) such as dwell time, traffic congestion, accidents, and so on. They usually prefer to minimize the time spent traveling from their origin to destination by choosing the best route.

On the other hand, travel time variability comes from various sources, which can be divided into two categories: regular variations (recurrent) like e.g. day-to-day variation and irregular condition variations (non-recurrent) like e.g. incidents, weather or random variations [48]. For non-recurrent variations, it is hard to predict the location and time of their occurrence [43]. Therefore, it is hard to predict the bus travel time that would result for passengers from adjusting their departure time in such cases [78], [28].

By contrast, with known regular and irregular condition-dependent variations, travelers may be able to adjust their departure time or route to arrive on time at their destinations [78].

In recent years, Intelligent Transportation Systems (ITSs) has been widely used in planning, evaluation and control of the reliability of public transportation systems [39]. Most public transportation systems such as bus services have either successfully implemented or are in the process of implementing various ITSs applications in their system, with the aim of providing reliable and accurate information for passengers [89], [19], [33].

Over the past several years a number of research projects have attempted to empirically measure behavioral responses to changes in travel time variability [84]. These have generally been built on theoretical models of scheduling choice that account for changes in departure time in response to the expected punctuality associated with variability. Using the mean variance model approach, these studies have confirmed that the travel time tends to be influenced by traffic conditions, ridership and weather conditions, which, in turn, may present variability depending on the time periods in the day and the day of the week [32], [27].

On the other hand, an important evolution in Intelligent Transportation Systems (ITSs) is largely due to the availability of bus probe data. Probe data generated by vehicles include data obtainable from navigation systems, such as the time and position (longitude and latitude), i.e. data on the vehicle's running and performance history [69], [77], [86]. Since these probe data can be obtained continuously over time from a vehicle, they enable monitoring of the state of road traffic characteristics [77]. It is expected that detailed traffic analysis of bus travel times could be carried out using these data before making a prediction model.

1.2 Problem Formulation

In recent years, many studies have been performed to develop models to predict bus travel times on routes, especially arrival times at bus stops. In addition, a number of dynamical models have been proposed to predict bus travel time on roads in urban areas. However, such models are over-simplified and may not represent real conditions because the models do not take into account bus travel time variability on the routes [27], [77].

In summary, previous studies have been conducted in the research field of predicting bus travel/arrival times for a single bus route using the historical average bus travel/arrival time. Furthermore, their prediction models assume that the historical travel time patterns will remain the same even in the future time period in a day . In this case, model precision is highly dependent on the amount of the historical traffic pattern data, as the accuracy of analysis results may differ depending on real conditions [30], [63].

In addition, limitations in the volume of historical data make for a significant difference in the relationship between historical traffic patterns and real conditions. The problem in these methods usually comes from the assumption that travel time recurs predictably or the assumption that a regular pattern repeatedly occurs in the same time-period over days [22]. Therefore, the variations in bus travel time on the route are often not sufficiently well-defined to build a prediction model.

Liu et al. [52] noticed that a bus is not operated in the same way as other vehicles. Even though the bus delay is caused by a traffic jam, the bus cannot increase its speed to adjust the delay because the bus speed is limited and they should follow the route that has been determined by a time schedule [52], [16]. Many authors also noticed that delays at the upcoming bus stops depend on accrued variation of travel time at past stops. Bertini et al. [30] showed that the bus travel time between two adjacent bus stops increases in

several time periods. Moreover, there are correlations of travel time between time periods, especially two adjacent time periods, which influence the bus travel time in the next time period [55], [77], [71].

In other words, the bus travel time in the previous period will affect the travel time in the next or later period [23]. Likewise, if the trend of the historical travel time is not linear, e.g. when the short-term fluctuations due to accidents that happen suddenly in the current period will affect later travel time periods such as the morning peak period, variations in the historical dataset and variations in the relationship between the historical patterns and the current traffic patterns could dramatically affect the prediction in a negative way [78]. Therefore, the performance of these models is highly dependent on the quality of the historical data.

1.3 Uniqueness of this Research

Numerous studies have been conducted to predict the bus travel time between two adjacent bus stops considering the variability of travel time between time periods in a day. However, these models do not give sufficient attention to predicting bus travel time over intervals between adjacent bus stops. The result is that these models do not have the ability to capture the complex non-linear relationship between travel time and variability.

Further, in this thesis, I propose a model for predicting travel time over unstable intervals between adjacent bus stops on the routes during different time periods in a day, since bus travel time over the unstable intervals varies significantly between time periods in a day and in the same time period over days, and there are in addition strong correlations between time periods in a day. In constructing my model, the variability of travel time is the main focus point in the prediction model.

Therefore, I consider using two important independent variables: historical average travel time data in the same time period and dynamic average travel time data in the time period just before the current one. Using these independent variables yields a highly accurate performance in dynamically predicting bus travel time under recurrent and non-recurrent variability. Therefore, the model may capture the influence of unexpected events such as accidents, and traffic jams that influence the bus travel time on the routes.

1.4 Research Objective and Scope

The primary objective of this study is to develop a nonlinear dynamical model for predicting bus travel time over unstable intervals between adjacent bus stops in seven and eight time periods in a day. This study uses real bus probe data to develop the model. The study mainly examines the significant factors: variations of travel time between time periods in a day and in the same time periods over days as well as the correlation of travel time between eight time periods in a day, which influences the bus travel time in the current time period. In general, the objectives of this research are outlined as follows:

1. Calculate bus travel time over each interval between two adjacent bus stops in each time period, where I divide a day into eight time periods: early morning (EM), morning peak (MP), late morning (LM), midday (MD), early afternoon (EA), afternoon Peak (AP), evening (E), and late night (LN). In Section 5, the early morning (EM) period is ignored and only the remaining seven periods are considered.

2. Clarify that the bus travel time shows significant variation depending on the time period and the day.

3. Classify intervals between two adjacent bus stops into two classes: stable and unstable.

4. Establish the variability of bus travel time over unstable intervals between the eight time periods in a day and that in the same time period over days using a statistical test.

5. Provide the correlation of bus travel time over unstable intervals between time periods in a day using a statistical test.

6. Build a model to predict bus travel time over each unstable interval in each of seven and eight time periods in a day. Then, evaluate the prediction model by comparing them with other models.

1.5 Research Framework

Figure 1.1 shows the conceptual framework of this research for predicting bus travel time over each unstable interval. I defined the framework of my study as follows:

1. Data Preparation and Preliminary Analysis: The process begins with the preparation and preliminary analysis of the probe data. I extract the information from bus probe data and observe bus travel time over each interval between two adjacent bus stops in eight time periods in a day over 20 days. This is to show that bus travel times are influenced by each time period, in which traffic conditions, ridership and weather conditions often change even though buses run over the same intervals between adjacent bus stops. Then, using the standard deviation of travel time in each time period, I roughly classified all intervals into two classes: stable and unstable. The details are described in Chapters 3 and 4.

2. Advanced Analysis: This stage covers the variability analysis of travel time over each unstable interval for model development. I conduct three statistical analyses to confirm the variations of the travel time among eight time periods in a day, in the same time periods over days and the correlation of the travel time between adjacent time periods in a day. The details are described in Chapter 4.

3. Building of Prediction Models: Considering the results obtained in stage 2, I chose two significant factors influencing the bus travel time over each unstable interval as input parameters of the models. I conducted experiments to predict travel time over unstable intervals focusing on recurrent and non-recurrent variability between the seven time periods in a day. To build the prediction models, I applied a time series approach using three machine learning methods: Artificial Neural Network (ANN), Support Vector Machine (SVM) and Random Forest (RF). The details are described in Chapter 5.

4. Establishment of prediction models using another approach which distinguishes off-peak and peak-hour periods: In this stage, I conducted experiments to predict travel time over unstable intervals focusing on off-peak and peak-hour periods for eight time periods in a day. To build the prediction models, I applied the time series approach using two machine learning (ML) techniques: Artificial Neural Network (ANN) and Support Vector Machine (SVM). The details are described in Chapter 6.

5. Evaluation and Comparison: I compare the model's performance using a wide range of different types of data sets to decide which are the most suitable input variables. Then, I identified the influence of different attributes in the input variables which have a significant effect in the prediction results. The details are described in Chapter 7.

1.6 Research Question

The major research questions of our work concern revealing the factors involved in and ways to achieve high-accuracy prediction results from real bus probe data with different parameters of independent variables. There are many parameters that will affect the accuracy of results to be considered before building a prediction model. The following are the research questions addressed in this research.

- Question 1: Are there any differences between travel time over each interval between two adjacent bus stops over the eight time periods in a day?
- Question 2: Is it possible to set boundaries between stable and unstable intervals?
- Question 3: What is the ratio of the unstable intervals to the whole?
- Question 4: Are there any variations in unstable intervals among the eight time periods in a day and among the same time periods over days?
- Question 5: Is there any correlation of travel times over unstable intervals among the eight time periods in a day?
- Question 6: Is it possible using nonlinear dynamical models built using machine learning techniques such as ANN, SVM and RF to predict bus travel time over each unstable interval between two adjacent bus stops?
- Question 7: Are there any significant impacts in using different input variables to predict travel time over each unstable interval while distinguishing between off-peak and peak-hour periods?

1.7 Contributions

Three contributions have been made in this paper. First, the records of bus travel time over each interval between adjacent bus stops were obtained from all of the bus routes during eight time periods in a day over 20 days. Then, the intervals were distinguished into stable and unstable intervals.

Second, I clarified the variability of bus travel times over unstable intervals between time periods and in the same time periods over days. I also identified that there are statistically significant correlations of travel time between the eight time periods in a day which influence the bus travel time in the current time period over unstable intervals. Third, I developed nonlinear dynamical models to predict bus travel time over each unstable interval between adjacent bus stops. The characteristics of the models are as follows:

1. A prediction model to predict travel time in each of the seven time periods in a day focused on regular variations (recurrent) and irregular condition variations (non-recurrent).

2. A prediction model to predict the bus travel time in each of the eight time periods in a day based on traffic density i.e., off-peak and peak-hour periods. In this model, I demonstrated the impact of two types of input variables for the prediction in off-peak and peak-hour periods.

Finally, to measure the prediction performance, I evaluated the performance of the models by conducting several experiments. First, I conducted a comparison experiment between our proposed model and the model in other previous study. Second, I compared the proposed model and the model in my previous study.

1.8 Thesis Outline

The thesis is organized as follows: Chapter 1 discusses the background, problem formulation, uniqueness, objectives, framework, questions addressed and contribution of this research. Chapter 2 presents a literature review of conceptual, theoretical and methodological topics related to travel time variability and prediction models for bus travel time. Chapter 3 explains the data used in this research, shows how real-world bus probe data can be calculated/utilized and introduces briefly several machine learning techniques used. Chapter 4 describes an empirical analysis of the distribution the bus travel times, and shows statistical analyzes to confirm the variations of travel time over each unstable interval. Chapter 5 presents the simulations of the proposed model's algorithms to predict bus travel time focusing on recurrent and non-recurrent variabilities of travel time over the unstable intervals. Chapter 6 provides a different approach for predicting bus travel time considering traffic congestion by distinguishing peak periods from off-peak periods. Chapter 7 presents the model performance evaluation resulting from the comparison experiment between our models and the model proposed by other study , as well as the model in my previous study. In Chapter 8, the thesis concludes with a summary of the results of the experiments and the findings and contributions of this research , and discusses some topics for future scientific research in the prediction of bus travel time.

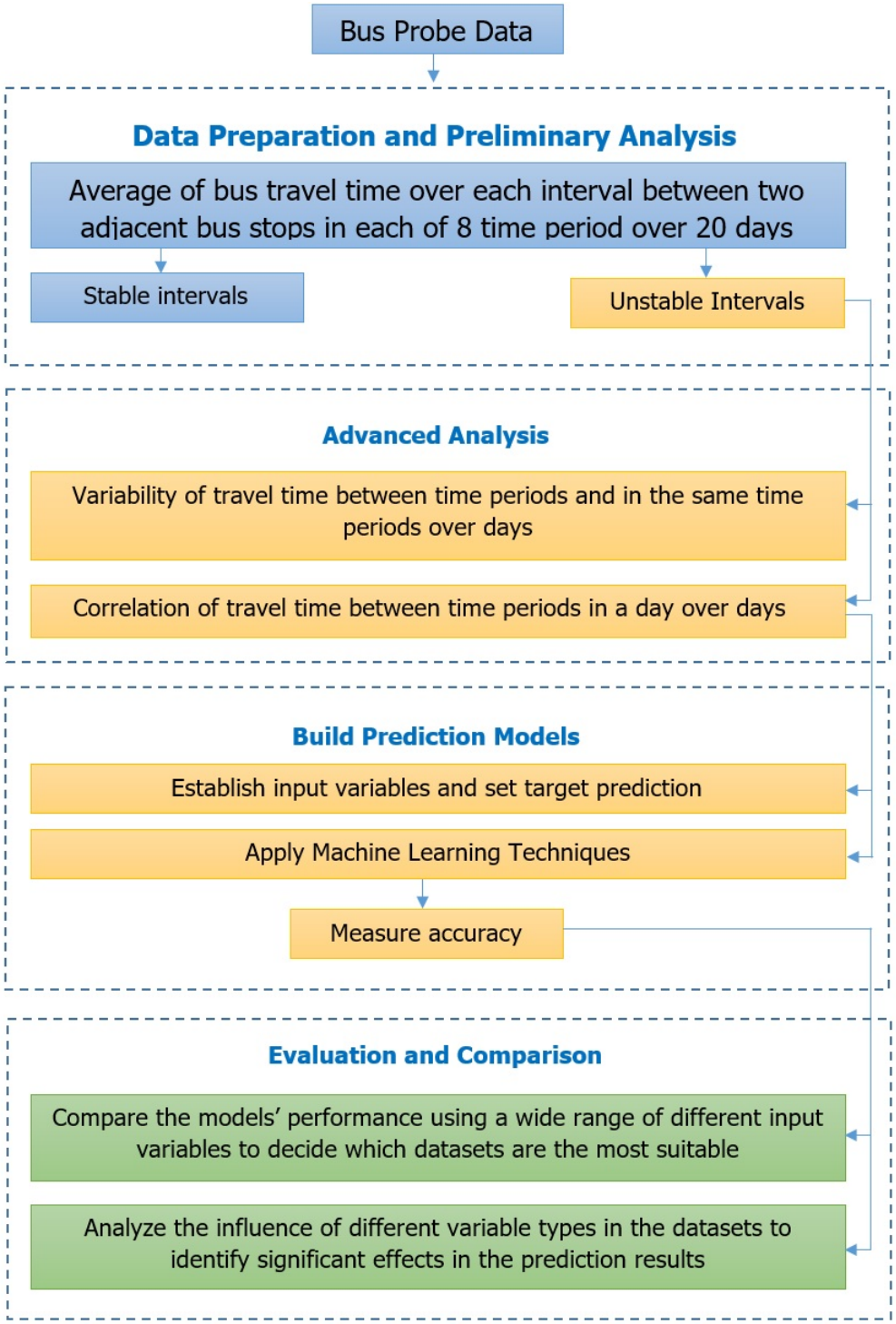

FIGURE 1.1: Framework of this research

Chapter 2

Related Work

2.1 Introduction

This chapter presents a literature review of conceptual, theoretical and methodological topics related to the prediction of bus travel and arrival time. Through the literature review, the importance of models for the prediction of bus travel/arrival times for passengers and management control of bus travel time became clear. The literature review led to the motivation to test a stochastic time series in nonlinear dynamical models for bus travel prediction.

2.2 Travel Time Variability

Travel time variability reflects the degree of variation in the travel time of a trip that is recurrent or non-recurrent over several time periods in a day or day to day [73]. Traffic congestion as a source of travel time variability should be analyzed by distinguishing recurrent congestion (e.g., the daily increase in traffic during the peak hour periods on weekdays) and non-recurrent congestion, which is caused by infrequent incidents such as accidents and extreme weather [22], [43].

Travel time variability is a key factor that passengers consider when making basic travel decisions regarding destination, route and departure time,

and numerous studies have attempted to measure travel time variability using two modeling approaches: the scheduling model and the mean variance model [3], [9], [12], [43]. Measuring travel time variability enables greater support for the prediciton of travel time either with empirical, analytic or simulation methods [9]. Over the last several years many studies have investigated bus travel time variability on routes in urban networks, especially between adjacent bus stops along the route. Using the mean variance model approach, these studies confirmed that travel time tends to be influenced by traffic conditions, ridership and weather conditions, which, in turn, may vary depending on time period in a day and the day of the week [3], [9], [78].

Uno et al. [77] proposed a methodology for evaluating the road network from the viewpoint of travel time stability and reliability using bus probe data. Travel time distributions of arbitrary routes are estimated by statistically and directly summing up observed multiple travel time distributions. In their study, probe data can be applied to the tasks of automatic incident detection and observations of travel time and its variability.

In addtion, Gurmu [35] and E Durán-Hormazábal [43] conducted an analysis of bus travel time variability before building a prediction model. The model demonstrated its superior performance in terms of mean absolute percentage error (MAPE). Patnaik et al. [63] carried out an analysis of travel time variability over the eight time periods in a day. Their model could also predict bus arrival times for various conditions.

2.3 Methodology of Travel Time Prediction

A variety of models and algorithms have been developed to predict bus arrival times or bus travel times. The most widely used models can be classified into the following categories: historical average models, regression/time series models, Machine Learning Techniques (ML) including Artificial Neural

Network (ANN), Support Vector Regression(SVR) and Random Forest (RF).

2.3.1 Historical Average Model with Time Series Approach

Over the last several years, many researchers have empirically attempted to predict travel time over a route using historical average travel time directly or without combination with other inputs [7], [74], [71]. Historical average models are based on the historical data and able to predict the bus travel time or bus arrival time from previous bus trips. These models will be practical, useful, and reliable. Gurmu and Nall [9] developed a historical data model for predicting the link travel time between two bus stops, which was calculated as the average travel time between two bus stops minus the average dwell time at the bus stops. Patnaik et al. [43] also suggested a historical approach in their study and showed good results.

The strength of historical time series data models is high computation speed due to the simple formulation of the algorithm. The models do not need a large number of travel time variables, but only time-related data [85], [44], [92]. The models could be built only from historical data, without dynamic observations [18]. However, the main disadvantage of this type of model is the averaging of input data over time. The predictions of travel time tend to concentrate on the trend of the historical travel time data and become problematic if the trend is not linear, e.g. if short-term fluctuations due to an accident that happened suddenly in the early morning affect later travel time such as in the morning peak [28], [50]. Variations in the historical dataset and variations in the relationship between the historical patterns and the current traffic patterns could dramatically affect the prediction in a negative way [41]. Moreover, the performance of these models is highly dependent on the quality of the historical time series dataset, which is not always available [19], [92], [62].

Furthermore, a lot of proposed studies also discuss historical average travel time using regression models. In addition, their models require a linear mathematical function to explain a dependent variable with a set of independent variables [90]. Unlike the previous models, these are able to work satisfactorily even if traffic conditions are not stable. They usually measure the simultaneous impact of various factors, which are independent of one another, affecting the dependent variable. For example, Patnaik et al. [63] developed a set of multiple linear regression models to estimate bus arrival times using distance, number of stops, dwell times, number of boarding and alighting passengers and weather descriptors as independent variables. Their study showed that the models could be used to estimate bus arrival/travel time at downstream stops.

Jeong and Rilett [40] and Ramakrishna et al. [65] also developed multiple linear regression models using different sets of independent variables. In their studies, the regression models outperformed the time series model. However, these models have a relative advantage in revealing which independent variables are less or more important for predicting travel times. For example, Patnaik et al. [63] mentioned that weather was not an important input in their model. Ramakrishna et al. [65] also found that two variables, i.e. bus stop dwell times from the origin of the route to the current bus stop in minutes and intersection delays from the origin of the route to the current bus stop in minutes, are less important in predicting bus travel time. Because variables in bus travel time are inter-correlated between time periods, the applicability of the regression models is in general limited [21], [74], [16].

On the other hand, machine learning techniques have recently gained popularity in predicting bus arrival/travel time. These techniques have also been used in several large-scale prediction competitions and suggest that by combining the model with the time series approach, the prediction accuracy can often be improved [47], [9].

2.3.2 Machine Learning Techniques

Machine learning (ML), which is a branch of artificial intelligence, is about the construction and study of systems that can learn from data. ML methods consist of two stages, i.e., choosing a candidate model, and next, predicting the parameters of the model through a learning process based on existing data [29]. ML methods have certain benefits with respect to statistical methods in the following respects: dealing with complex relationships between predictors that can come up within a huge volume of information, processing non-linear relationships between predictors, and processing complicated and noisy data. These models can be used for prediction of travel time, without implicitly addressing the traffic data [47], [9], [29]. Results obtained for one location are normally not transferable to the next, because of location-specific circumstances, e.g., geometry or traffic control.

In recent years, machine learning techniques (ML) have commonly been used to predict travel time because of their ability to solve complex non-linear relationships. ANN was demonstrated as a potential method for predicting travel time. Chung, E. H., and Shalaby, A. [71], Bai et al. [9], Gurmu et al. [35], [29] developed models to predict bus arrival time with a variety of traffic conditions and introduced an ANN model based on historical data such as Automatic Vehicle Location (AVL), Automatic Passenger Ticket Issue System (APTIS) and GPS data. Their proposed models are suitable for finding complex nonlinear relationships between the dependent variable of bus travel time and the independent variables that influence the travel time [11], [42]. Moreover, these are data-driven techniques and require a large set of data for better learning. Also, they are problem-specific models and whenever the input variables change, the whole model has to be restructured [83].

On the other hand, a lot of studies [2], [9], [19], [34], [85], have employed other machine learning techniques such as Support Vector Machine (SVM)

and Random Forest (RF) to build prediction models for bus travel time and showed that these models were practical, useful and reliable, where the traffic flow was relatively small and stable.

In addition, RF has also been applied to prediction of bus travel time under traffic flows and showed that the model outperformed other models in terms of prediction precision [34], [58] [36], [46]. Although their work uses bus historical data with a consideration of traffic conditions and with theday divided into several time periods, their model only focuses on a number of routes in certain corridors.

In order to construct a prediction model using real travel time data (historical), previous studies employed a combination of ML and time series model. Next, their model explicitly incorporated information about seasonality into the data (time period of a day and day of a week, etc) using bus probe data [5], [29], [35], [80]. The models confirmed that the developed model could be applied to predict short-term travel time with various conditions. However, they have not discussed a model for predicting travel time over each unstable interval between adjacent bus stops considering the variability of travel time over time periods and days.

Moreover, the literature survey shows that most reported studies about bus travel time/arrival time prediction have been developed for homogeneous traffic conditions only. This is because heterogeneous traffic conditions are very complex and even their analysis to build a prediction model may be more challenging [8], [61], [64].

2.4 Summary

The above literature review of the models and algorithms for bus travel time prediction shows that many models are based on historical patterns and other variables correlated with the arrival/travel time. The variables used

include real data about historical arrival or travel time, dwell time, number of stops along the route, distance between adjacent stops and the road characteristics. They are from data collections such as, AVL, APTIS, survey and Probe data.

History-based models assume that the conditions of traffic do not change much, which may not be true when considering a switch from off-peak to peak-hour and vice versa. These models were mainly used in areas where congestion is minimal because they assumed traffic conditions are similar (homogeneous). However, it could be argued that it is also possible to observe such patterns in areas where the congestion is severe. This can be found out from extensive historical data analysis by looking into the distribution of travel time between time periods over days or days of the week and so on. In areas where stable demand and similar traffic patterns exist, history-based models are able to give satisfactory bus travel time information. So there is no need to go for complex prediction models. Machine learning techniques such as ANN, SVR and RF have outperformed other methods in cases where enough data is available.

However, to greatly improve the prediction accuracy of travel time, we should first focus on shorter intervals on a route, such as the intervals between adjacent bus stops, considering the variability of travel time between them. Furthermore, it is difficult in practice to determine whether or not a time series of travel time recurs and whether a regular pattern repeatedly occurring in the same time period over days becomes a homogeneous or heterogeneous traffic pattern. This heterogeneity, coupled with variability of travel time on the road makes bus travel time prediction more challenging than can be handled by the reported proposed model. There is, therefore, a need for models that can capture the stochastic behavior of traffic characteristics with a large data requirement.

The present study will be an attempt in this direction to develop nonlinear

dynamical models for predicting bus travel time. For this purpose, at first, I classified intervals between two adjacent bus stops into two classes: stable and unstable. Next, I identified two statistically significant factors: variations of travel time in the same time periods over days and the correlation of travel time between the seven or eight time periods, which influences the bus travel time in the current time period over unstable intervals. Then, I developed nonlinear dynamical models for predicting bus travel time over each unstable interval between adjacent bus stops in each of seven or eight time periods in a day.

Chapter 3

Data and Methodologies

3.1 Introduction

This chapter explains the data used in this research, describes the preliminary data preparation for calculation and the methods of prediction models. The present study shows how real-world bus probe data can be utilized, how the data should be calculated, how to distinguish stable intervals from unstable intervals and introduces briefly several machine learning techniques used.

3.2 Probe Data

3.2.1 Overview

Probe data generated by vehicles includes data obtainable from navigation systems, such as the time and position (longitude and latitude), i.e. data on the vehicle's running history, and front-rear acceleration or right-left acceleration, i.e. data on the vehicle's performance history. Since these probe data can be obtained continuously over time from the vehicle, they allow monitoring of the state of road traffic at any chosen location or point in time and the detection of traction information. Thus probe data offers the potential to develop a prediction model that can improve the accuracy of prediction results [6], [69], [77].

A typical unit of travel time for a bus on a route is the time required to move from one bus stop to the next, as shown in Figure 3.1, when a bus arrives at and departs from adjacent bus stops along route. A bus will travel on the same segment twice in a single trip [70], [45], but in the opposite direction i.e, the inbound and outbound directions.

3.2.2 Bus Probe Data

The probe data used in this research were provided by NISHITETSU Bus Company. The data were collected from the 21st of November to the 20th of December 2013. The probe data include GPS information on bus positions, time information indicating when the GPS information was taken, route number, number of bus stops, travel direction and so on. Bus routes are operated for around 18 hours a day. Buses run in different patterns on weekdays, Saturdays and Sundays/holidays according to their time tables.

TABLE 3.1: Fields of bus probe data

Field	Description
Vehicle Id	Bus identity number
Lat & long	GPS Position of Latitude and Longitude
Direction	Inbound or outbound
GPS time	Time obtained by GPS
Route Number	The number assigned to a route
Type of Route	Sub route number
Bus number	The order of a bus running on a route
Bus stop code	A code number assigned to a bus stop
Bus pole code	A code number assigned to a pole of a bus stop

Here, in this study, I have just dealt with travel time on weekdays, not Saturdays or Sundays/holidays because of the lack of data volume. I analyzed 175 routes consisting of 6129 intervals for the inbound direction and 5700 intervals for the outbound direction, on which 2045 buses a day were operated. Travel time information is recorded every 3 minutes and at the

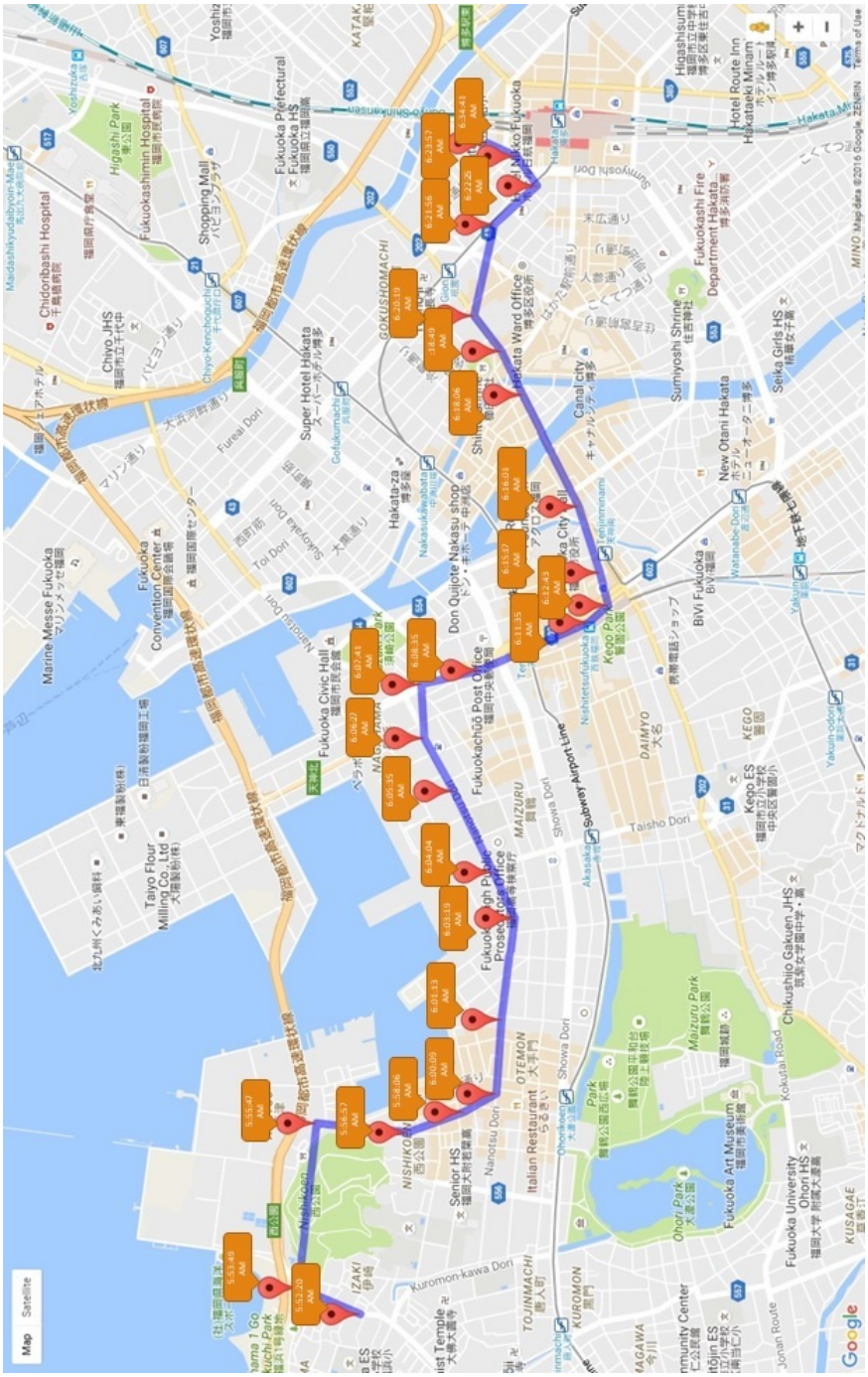

FIGURE 3.1: An example of one route where the bus runs between two adjacent bus stops

time when a bus stops. Table 3.1 shows a summary of bus probe data used in this study. In addition, the bus trips are classified in the database into eight different time periods, as shown in Table 3.2.

3.3 Methodologies

3.3.1 Travel Time over Intervals

In order to calculate bus travel time, first I calculated the average travel times over each interval during each of the eight time periods in a day. This is because bus travel times are influenced by each time period, in which traffic conditions, ridership and weather conditions often change even though the buses run over the same intervals between adjacent bus stops [3], [35], [63], [86].

Then, I calculated the average travel time over each interval in each time period in a day for 20 days, because the travel time may usually vary during the day. For short, I will just use the term "interval" below, when I mean "travel time over interval between adjacent bus stops", where the classification and definition of the time periods are shown in Table 3.2 and a bus travel time interval is illustrated in Figure 3.2.

TABLE 3.2: Time Periods

Periods	Ranges of Time
Early Morning (EM)	5:00:00-7:29:59
Morning Peak (MP)	7:30:00-9:29:59
Late Morning (LM)	9:30:00-11:59:59
Midday (MD)	12:00:00-12:59:59
Early Afternoon (EA)	13:00:00-15:29:59
Afternoon Peak (AP)	15:30:00-17:29:59
Evening (E)	17:30:00-19:29:59
Late Night (LN)	19:30:00-25:59:59.

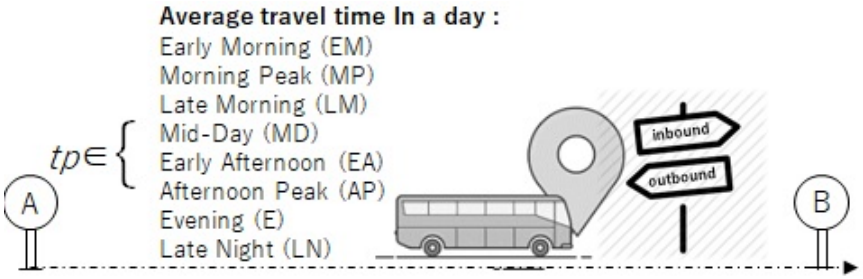

FIGURE 3.2: Interval between two adjacent bus stops

The following shows how to calculate bus travel time over an interval between adjacent bus stops. I define travel time $Tt_{AB}(i, tp, d)$, which is the length of time when bus #i runs between adjacent bus stops: A and B, in time period tp ($\in$ (EM, MP, LM, MD, EA, AP, E, LN)) on a day d, which is always a weekday in this paper,

as follows:

$$Tt_{AB}(i, tp, d) = t_B(i, tp, d) - t_A(i, tp, d), \tag{3.1}$$

where $t_B(i, tp, d)$ and $t_A(i, tp, d)$ are the time when the bus #i arrives at bus stop B and departs from bus stop A in time period (tp) on a day (d), respectively. Using equation (3.2), we calculate the average travel time $Tt(tp)$ in each of 8 time periods (tp) in a day as follows:

$$Tt_{AB}(tp, d) = \frac{1}{N} \sum_{i=1}^{N} Tt_{AB}(i, tp, d), \tag{3.2}$$

where N is the number of buses running on the interval between adjacent bus stops A and B, and may vary according to each interval in the specific time period. In what follows, we refer to each average travel time between adjacent bus stops for each of the eight time periods in day as an interval.

3.3.2 Distinguishing Stable and Unstable Intervals

The variability of bus travel time can be categorized by its time frame. Maria et al. [54] discussed variability as occuring between time periods in a day or between days. The variability is caused by unexpected events such as construction or inclement weather or generally refers to changes in travel time due to peak-hour congestion.

In addition to periodical and daily variations in travel time, it may be of interest to compare the average travel time between time periods in a day. Other studies suggest that travel times may vary for different time periods in a day due to changes in vehicle volume, construction etc. However, in the same time periods travel times should be similar for various days of the week in the absence of unexpected events. In this work, average travel time data for each of the eight time periods in a day are compared.

First, using equation (3.3), I calculate the average travel time $Tt_i(tp)$ over intervals i in time period tp over n weekdays, where $n = 20$.

$$Tt_i(tp) = \frac{1}{n} \sum_{d=1}^{n} Tt_i(tp, d) \tag{3.3}$$

Then, I put a number onto each interval from 1 to N, where N is the total number of intervals; I calculate Tt_i, which is the average travel time over intervals i ($1 \leq i \leq N$) among $TP = \{EM, MP, LM, MD, EA, AP, E, LN\}$, a set of eight time periods of a day using equation (3.4).

$$Tt_i = \frac{1}{|TP|} \sum_{tp \in TP} Tt_i(tp) \tag{3.4}$$

To transform all the data to normal distribution, I transform the data of the average travel time over intervals (Tt_i) among time periods (TP) in a day using natural logarithm.

Next, I calculate $StDev_i$, which is the standard deviation of the average

travel time over intervals i to find out whether travel time over each interval is stable or not. To make a fair comparison for all routes considering the differences of distance of all the intervals, I divide the standard deviation of the average travel time over each interval by the average travel time over the interval using equation (3.5).

$$StDev_i = \frac{\sqrt{\frac{1}{|TP|}\sum_{tp \in TP}(Tt_i(tp) - Tt_i)^2}}{Tt_i} \tag{3.5}$$

The second step is to create logarithmic ranges to distinguish criteria for stable and unstable intervals of travel time over each interval as shown in Figure 3.3. I calculate σ, the standard deviation of the standard deviation of the average travel time over all the intervals using equation (3.6).

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(StDev_i - \mu)^2}{N-1}} \tag{3.6}$$

Here μ in equation (3.7) is the average of the standard deviation over the intervals.

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(StDev_i) \tag{3.7}$$

Next, I calculate the standard deviations of the average travel time over interval i ($1 \leq i \leq N, where N$ is number of intervals) for each of the eight time periods in a day using equation (3.8)

$$StDev(Tt_i) = \sqrt{Var(Tt_i)} \tag{3.8}$$

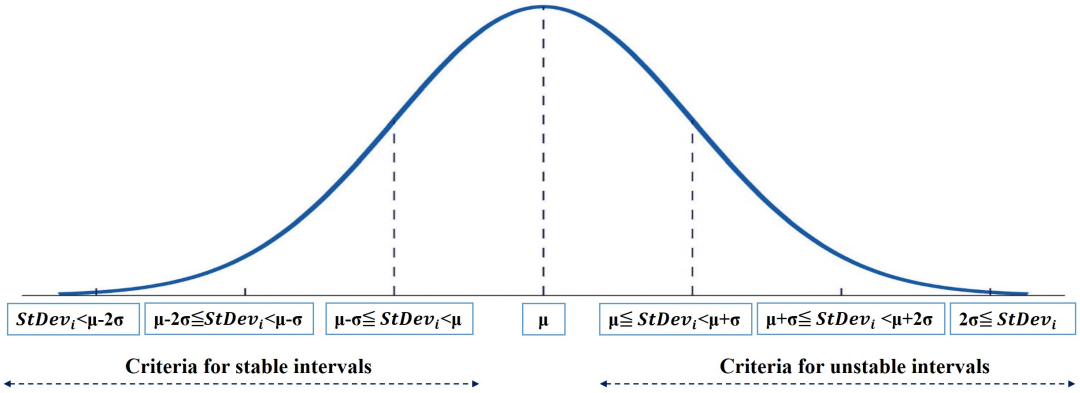

FIGURE 3.3: Logarithmic ranges stable and unstable intervals

According to $StDev_i$, interval i is classified into logarithm ranges of interval criteria as shown in Figure 3.3 and Table 3.3. I roughly classified all intervals into two classes: stable and unstable; if the standard deviation of travel time over an interval is less than μ, then the interval is classified as stable, or otherwise, unstable. Then, I further classified each of the two into three sub-classes: weak, medium and strong (we present the results in Section 4.4). The criteria for each subcategory are as follows:

TABLE 3.3: Logarithmic ranges of interval criteria

Interval Category	**Logarithmic Ranges**
Strong stable	if $StDev_i <= \mu - 2\sigma$
Medium Stable	if $\mu - 2\sigma < StDev_i <= \mu - \sigma$
Weak Stable	if $\mu - \sigma < StDev_i <= \mu - \sigma$
Weak Unstable	if $\mu < StDev_i < \mu + \sigma$
Medium Unstable	if $\mu + \sigma <= StDev_i < \mu + 2\sigma$
Strong Unstable	if $StDev_i >= \mu + 2\sigma$

3.4 Methods of Travel Time Prediction

In this Section, I briefly present references and an outline for nonlinear time series prediction using the machine learning techniques Artificial Neural Network with NARX, Support Vector Regression (SVR) and Random Forest (RF) Regression.

3.4.1 Time Series Approach

The purpose of this section is to provide a brief sketch of time series prediction theory. The accuracy of time series prediction is fundamental to many decision processes and hence research to improve the effectiveness of prediction models has been ongoing. Successful time series prediction is a major goal in many areas of travel time prediction. However, the time series data are often full of non-linearity and irregularity. There are vast amounts of technical references, books, and journal articles detailing time series prediction algorithms and theory for both linear and non-linear prediction applications [13], [68].

Fundamentally, "the goal of time series prediction is to estimate some future value based on current and past data samples" [68]. Mathematically, the prediction approach is stated as follows:

$$\hat{x}(t + \Delta_t) = f(x(t - a), x(t - b), x(t - c)), \tag{3.9}$$

where, in this specific example, $\hat{x}$ is the predicted value of a (one dimensional) discrete time series x. "The objective of time series prediction is to find a function $f(x)$ such that $\hat{x}$, the predicted value of the time series at a future point in time is *unbiased* and *consistent*. Where i is an index to a discrete time series value and N is the total number of samples. It should be noted that another measure of a predictor's goodness is efficiency as related to bias. If the

estimator achieves this bound, then it is said to be efficient" [38]. Estimators generally fall into two categories: linear and nonlinear. Over the past several decades, a vast amount of technical literature has been written about linear prediction: "the estimation of a future value based on the linear combination of past and present values" [68]. Real-world time series prediction applications generally do not fall into the category of linear prediction. Instead, they are typically characterized by non-linear models. Therefore, most non-linear models can be handled by machine learning techniques such as, ANN, SVM and RF [13], [68].

3.4.2 Artificial Neural Network (ANN)

I use an ANN-based time series prediction method to predict travel time over intervals between adjacent bus stops on all of the routes. The method is based on a Nonlinear Auto Regressive model with the eXogenous input (NARX) model. The NARX model is well-suited for modeling dynamic nonlinear systems, especially those with time series characteristics. In addition, "NARX model is a subset of the Nonlinear Auto-Regressive Moving Average with Exogenous Inputs (NARMAX), which are nonlinear non-parametric identification models" [26], [87].

The mathematical function which models a real-world system is very complex and usually unknown. However, the NARX model can be constructed using a simpler function structure such as neural networks [20]. The NARX model formulation [20], [26] is described as follows:

$$y(t) = f(y(t-1), ..., y(t-n_y), u(t-1), ..., u(t-n_u)) + e(t) \tag{3.10}$$

, where $y(t)$, $u(t)$ and $e(t)$ are the model output, model input, and noise at time t, respectively. n_y and n_u are the maximum lags in the output and the input, respectively; $f(.)$ is some vector-valued non-linear function, but can be

approximated using some known simpler function such as neural networks [20].

In my model, I used a multilayer perceptron (MLP) with a single hidden layer to approximate any bounded continuous function. The MLP contains one or more layers of hidden units. "The hidden units enable the MLP to learn complex tasks and meaningful features from the input/output relationships" [57]. Moreover, "high degree of connectivity between the MLP layers is determined by the weights of the network" [87]. I conducted MLP training with the Levenberg-Marquardt algorithm and evaluated the model using the measure of mean squared error (MSE) in training and testing. The MSE is a default indicator in training the ANN model. The ANN model with the smallest MSE value is considered to be the best model.

3.4.3 Support Vector Machine Regression (SVR)

"Support Vector Machine (SVM) has been developed to work on a non-linear problem by incorporating the concept of the kernel in high-dimensional space; SVR is an application of SVM to the case of regression" [81], which was designed to overcome the over-fitting and to yield a good performance [79], [85].

I assume there are n numbers of training data (x_i, y_i) $(i = 1, \ldots, n)$, where x_i is an input vector, and y_i is a scalar output. With SVR, I want to assign a function $f(x)$, which has the significant deviation ε from the actual target y_i for all training data. If the value of ε becomes equal or near to 0, a good regression model is obtained [85].

The main purpose of the SVR model is to construct a linear model in m-dimensional feature space which input x is mapped onto. Using mathematical notation, the linear model $f(x, w)$ is given below [59], [10]

$$f(x,w) = \sum_{j=1}^{n} w_j g_j(x) + b \tag{3.11}$$

, where w_j and $g_j(x)$ denote the jth weight and nonlinear transformation, respectively, and b is a bias. Next, prediction performance is measured by the loss function $L(y, f(x,w))$. SVR uses a new type of loss function called ε-insensitive loss function proposed by Vapnik [81]:

$$L_\varepsilon(y, f(x,w)) = \begin{cases} 0 & if|y - f(x,w)| \leq \varepsilon \\ |y - f(x,w)| - \varepsilon & otherwise \end{cases} \tag{3.12}$$

"SVR performs linear regression in the high-dimensional feature space using ε-insensitive loss, and at the same time, tries to reduce model complexity by minimizing $||w||^2$. This can be described by introducing (non-negative) slack variables $\zeta_i, \zeta_i^*(i = 1, \ldots, n)$, to measure the deviation of training data outside the ε-insensitive zone" [59]. Thus SVR is formulated as a minimization of the following function:

$$R_{(w,\zeta)} = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*) \tag{3.13}$$

subject to :

$$\begin{cases} y_i - f(x_i,w) \leq \varepsilon + \zeta_i^* \\ f(x_i,w) - y_i \leq \varepsilon + \zeta_i \\ \zeta_i, \zeta_i^* \geq 0, (i = 1, ..., n) \end{cases} \tag{3.14}$$

, where $(1/2)||w||^2$ is a weight vector norm, and C is a regularized constant determining the trade-off between the empirical error and the regularized term. ε is called the tube size of SVR and it is equivalent to the approximation accuracy placed on the training data points [10]. By introducing optimal constraints, this optimization problem can be transformed into a dual

problem whose solution is given by:

$$f(x) = \sum_{i=1}^{n_{sv}} (a_i - a_i^*).K(x_i, x) + b \tag{3.15}$$

subject to : $0 \leq a_i^* \leq C, 0 \leq a_i \leq C$

, where n_{sv} is the number of Support Vector (SVs) and K is a kernel function. "The kernel parameters should be carefully chosen as they implicitly define the structure of high dimensional features and thus controls the complexity of the final solution. However, generalization performance, here prediction accuracy, depends on a good setting for parameters C, ε, kernel parameters, and input values (x) of the training data" [59].

I selected Radial Basis Function (RBF) as the kernel function in this study. For measurement of performance (C) in the training process, I selected RBF network (λ, ε) at the minimum error as an SVR model [59].

3.4.4 Random Forest (RF)

The present Section is not intended to provide a detailed description of Random Forest (RF); the parameters are described in Section 5.3.

"Random Forest (RF) Regression is a regression technique that combines the performance of numerous Decision Tree (DT) algorithms to predict the value of a variable" [14]. Therefore, regression using RF can be implemented for time series prediction purposes. That is, when RF receives a $\mathbf{u}$ input vector, made up of the values of the different evidential feature analyses for a given training area, RF builds k numbers of regression trees and averages the results [91], [24], [51].

"Assumed that the $\underline{\mathbf{u}}$ is a random vector with k elements, the aim is to predict v by estimating the regression function:

$$m(\mathbf{u}) = E[\underline{v}|\underline{\mathbf{u}} = \mathbf{u}] \tag{3.16}$$

given fitting sample:

$$S_s = ((\mathbf{u}_1 v_1), ..., (\mathbf{u}_s, v_s)) \tag{3.17}$$

which are independent realizations of the random variable $(\underline{\mathbf{u}}\ \underline{v})$. Therefore, the aim is to construct an estimate m_s of the function m.

A random forest is a predictor constructed by growing M randomized regression trees. For the j-th tree in the family, the predicted value at $\mathbf{u}$ is denoted by $m_s(\mathbf{u}; \underline{\theta}_j, S_s)$, where $\underline{\theta}_1, ..., \underline{\theta}_M$ are independent random variables, distributed as $\underline{\theta}$ and independent of S_s. The random variable $\underline{\theta}$ is used to resample the fitting set prior to the growing of individual trees and to select the successive directions for splitting" [76]. The prediction is then given by the average of the predicted values of all trees. Before constructing each tree, the observations are randomly chosen from the elements of $\mathbf{u}$. These observations are used for growing the tree.

"To avoid the correlation of the different trees, RF increases the diversity of the trees by making them grow from different training data subsets created through a procedure called bagging. Hence, some data may be used more than once in the training, while other data might never be used. Thus, greater stability is achieved, as it makes it more robust when facing slight variations in input data and, at the same time, it increases prediction accuracy" [14], [82].

On the other hand, when RF makes a tree grow, it uses the best feature or split point within a subset of evidential features which has been selected randomly from the overall set of input evidential features. "Therefore, this can decrease the strength of every single tree, but it reduces the correlation between the trees, which reduces the generalization error" [14].

"Another characteristic of interest is that the trees of an RF classifier grow with no pruning, which makes them light, from a computational perspective" [67]. However, "The performance of the RF algorithm depends on the tuning

of its parameters and the variable selection" (also known as feature selection) [14], [76].

3.4.5 Measures of Model Performance and Prediction Results

The model parameter is evaluated with emphasis on Root Mean Squared Error (RMSE) in comparing the validation data set. The model with the smallest RMSE value is picked up as the best performing. Next, I select Mean Absolute Percentage Error (MAPE) as the measure of prediction accuracy and calculate them for all prediction results. RMSE and MAPE are defined in the following equations:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(t_{Actual}(i) - t_{Predicted}(i))^2} \tag{3.18}$$

$$MAPE = \sum_{i=1}^{N} \mid \frac{t_{Actual}(i) - t_{Predicted}(i)}{t_{Actual}(i)} \mid, \tag{3.19}$$

where $t_{Actual(i)}$ is the observed bus travel time over an interval in each time period; $t_{Predicted(i)}$ is the predicted bus travel time over the interval in the same time-period. N is the number of observation .

Chapter 4

Preliminaries to Model Development

4.1 Introduction

This chapter presents an empirical analysis of the distribution of the bus travel times.

The analysis is based on the bus probe data provided by Nishitetsu Bus Company Fukuoka, Japan. The analysis focuses on bus travel time, which has also been the topic for most of the available literature on modeling and valuing travel time reliability. First, the shape of the travel time distribution is modeled in a simple way, with relationships between the average travel time and each of eight time periods in a day over days. Then, considering the variability of bus travel time over intervals between adjacent bus stops, I divide the intervals into stable and unstable intervals.

Next, in this Chapter, I also clarify the variation of travel time over unstable intervals as a prediction target. I conduct three statistical analyses to confirm the variations of travel time over each unstable interval i.e., between time periods in a day, in the same time periods over days and the correlation of the travel time between adjacent time periods in a day over days.

In addition, one aim of these analyses is to provide discussion of the input

variables (independent variables) required to successfully develop nonlinear dynamical models for bus travel time prediction, especially for unstable intervals.

4.2 Travel Time Pattern Analysis

Traffic patterns can be typically classified as monthly, weekly, daily and periodical. The periodic pattern analysis checks whether the travel time data during different time periods in a day or in the same time periods day to day have a similar pattern or not. Similarly, travel time is compared between time periods in a day and the daily patterns are compared with those of the corresponding months, weeks and days [31], [37], [78]. In addition, the pattern analysis of travel time and comparison between time periods checks whether the current travel time has a pattern similar to the previous travel time on the same day. This is an early stage of the analysis of bus travel time during different time periods in a day over days and may help to capture traffic conditions on a particular day and related events such as accidents, congestion, etc. [1], [60].

For the purposes of this study, using the earlier method mentioned in Section 3.3, bus travel time series were obtained over each interval between adjacent bus stops for all routes in each of eight time periods in a day for 20 days. Next, I investigated the relationship between travel time over intervals between adjacent bus stops in each time period of all routes day to day and clarified the daily characteristics of bus travel time. I also intended to prove an assumption that travel time in a particular time period has a strong correlation with travel time in other time periods in a day; during off-peak hours such as early morning (EM), midday (MD), or late night (LN), both traffic volume and travel time decrease. Meanwhile, in the peak-hour such as morning

peak (MP), late morning (LM), and evening (E), the traffic volume increases dramatically and the travel time increases as well.

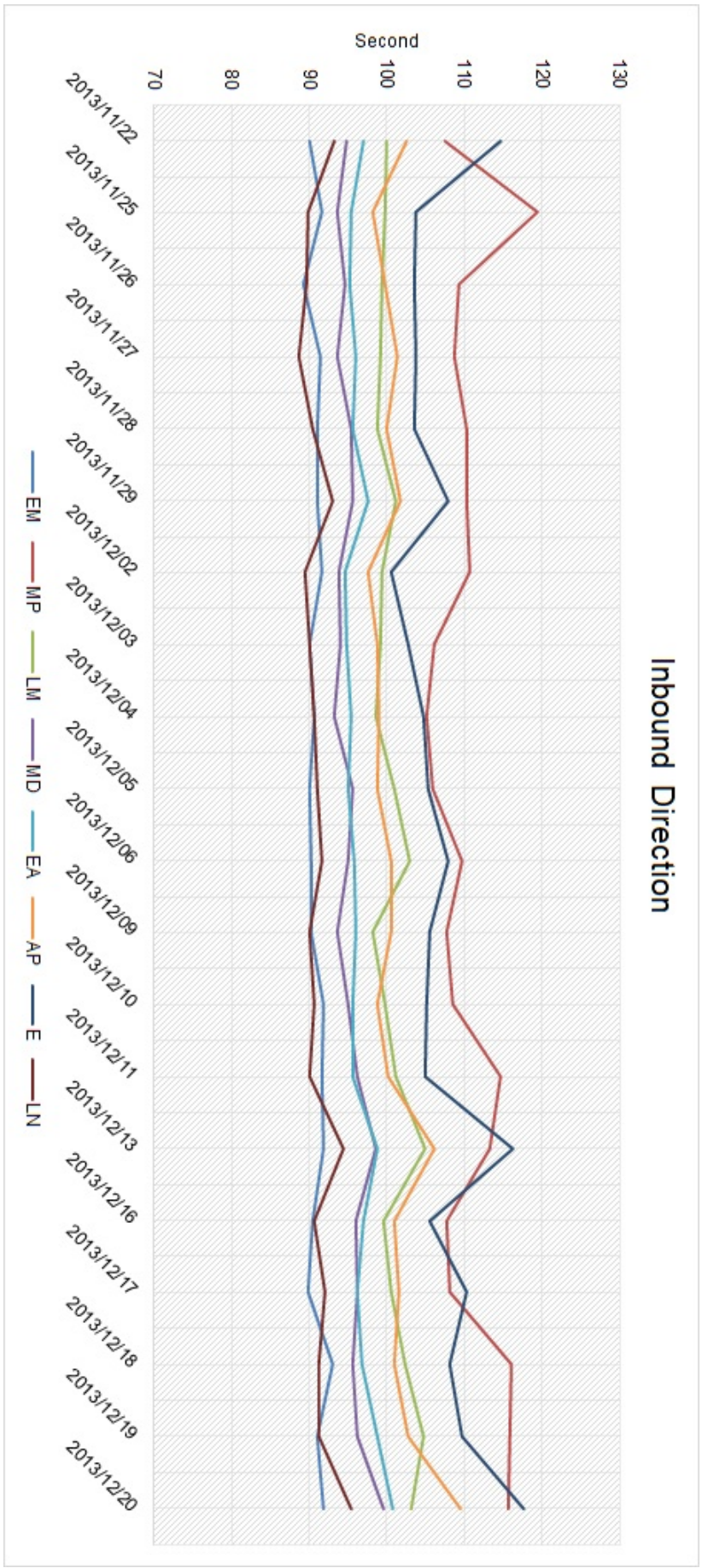

FIGURE 4.1: Daily average of bus travel time over intervals for the inbound direction

Figures 4.1 and 4.2 show a relatively constant travel time over each interval for both directions during early morning (EM), Midday (MD), early afternoon (EA) and late night (LN), indicating that travel time between adjacent bus stops tends to exhibit a homogeneous pattern (similar conditions) in the same time periods day to day. At the same time, in several time periods in both directions, such as morning peak (MP), late morning (LM), afternoon peak (AP) and evening (E), there are significant variation in travel time, and I assume that many people traveled during these time periods. These results explain why I chose the time period as of the appropriate level of analysis for calculating the average travel times for each day.

4.3 Empirical Study on the Travel Time Variability

The variability of bus travel time has often been described, and it can refer to changes in travel time during different time periods in a day or changes from day to day [54]. The variability between days is caused by unexpected events such as construction or weather and can be recurrent or non-recurrent. Meanwhile, the variability of travel time between time periods in a day, generally refers to changes in travel time due to congestion in peak-hour and usually tends to be recurrent [86], [85], [56]. Thus, the calculation of the travel time variability over each interval on each route has the advantage that the properties of the distribution can be recognized when the bus runs between adjacent bus stops in a day [88]. This is because travel time in several periods can be compared to find out whether the travel time over intervals varies or not [17] (e.g. with the assumption that, if travel time over each interval differs between time periods in a day, the interval is variable, as mentioned in Section 4.2).

Furthermore, I also investigated two topics: first, variability of bus travel time between time periods to distinguish stable and unstable intervals, and

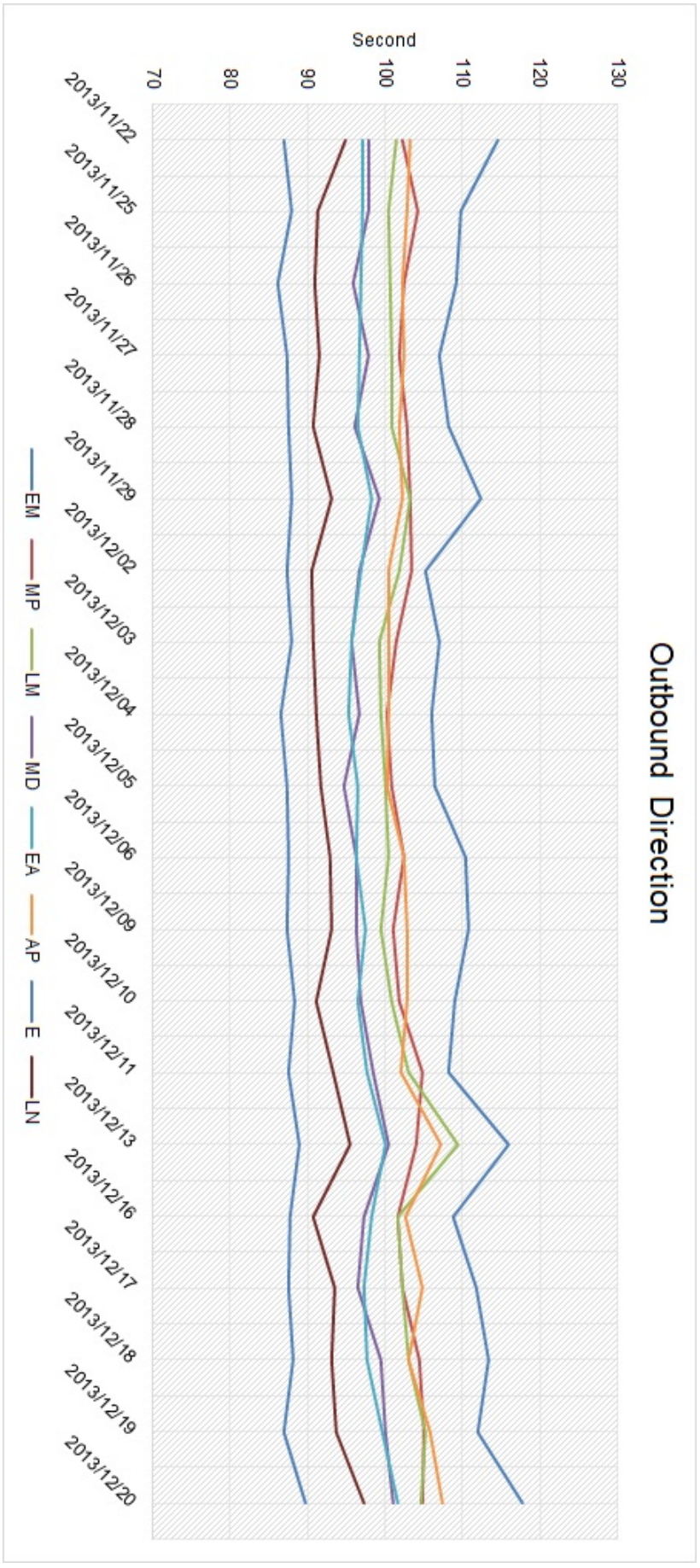

FIGURE 4.2: Daily average of bus travel time over intervals for the outbound direction

the variation in travel time over unstable intervals.

4.4 Stable and Unstable intervals

Bus travel time over intervals between adjacent bus stops during different time periods may vary due to traffic congestion and ridership situations along the route and therefore deviate between time periods in a day. For example, in the afternoon peak (AP), people are likely to use buses to do shopping or errands; thus, the buses may serve more stops. Also, most people going to or from work are using buses or private cars in the morning peak (MP) and afternoon peak (AP), which may lead to significant increases in bus travel time and cause traffic congestion [66]. On the other hand, in the early morning (EM) and late night (LN), bus travel time is likely to be least impacted by traffic congestion. These facts signify that the time period is a significant factor associated with bus travel times.

Using the methods in Section 3.3.1 and Section 3.3.2, first, I calculated the average travel time in each of the eight time periods in a day over each interval for 20 week days and transformed the data to normal distribution using natural logarithm. Then, I calculated the variance over each interval for the eight time periods in a day and calculated the standard deviations of the average travel time over each interval for the eight time periods in a day. Following that, I created logarithmic ranges to distinguish stable and unstable criteria for travel time over each interval. To create the ranges, I first calculate the average of all standard deviations over intervals and the standard deviation of all the standard deviations over all the intervals. This results in a rough classification of all intervals into two classes: stable and unstable, with a further division of each of the two classes into three subclasses: weak, medium and strong. The results are as shown in Figure 4.3.

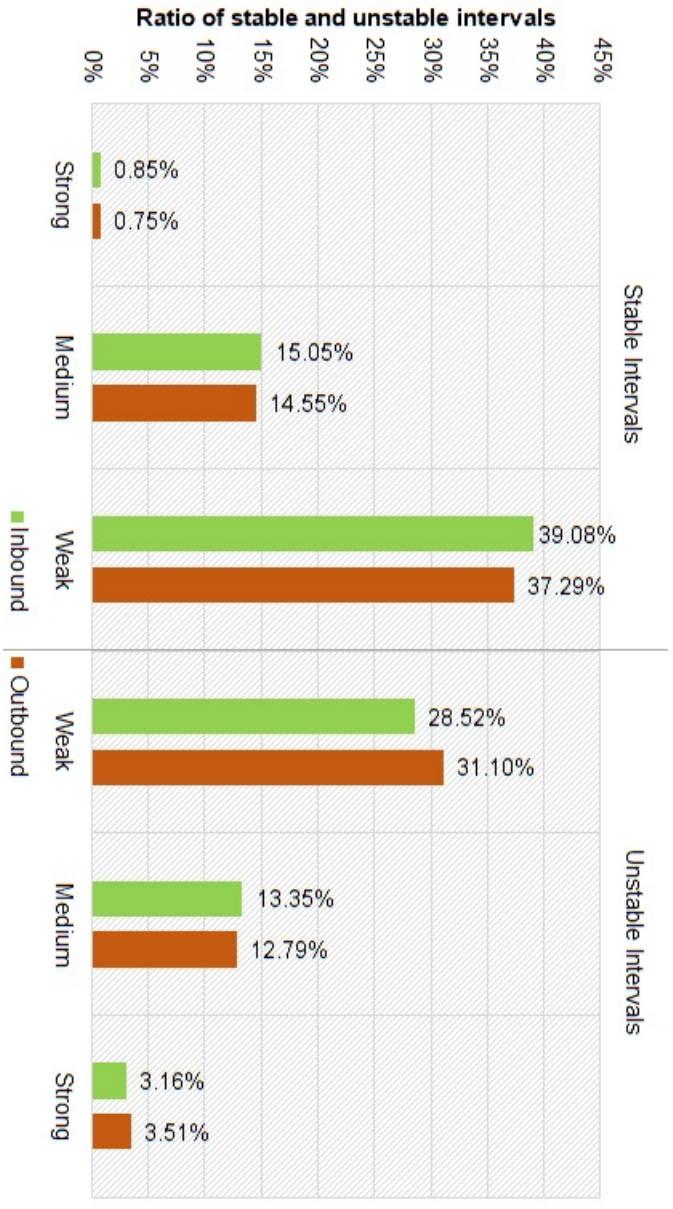

FIGURE 4.3: Ratio of stable and unstable Intervals

Figure 4.3 shows the percentage of stable and unstable intervals over days. There is a significant amount of variability in the travel time over each interval of all routes over days on weekdays. The percentages of three unstable interval classes: weak, medium and strong, are 28.52%, 13.35% and 3.16% for the inbound direction, and 31.1%, 12.79% and 3.51% for the outbound direction, respectively. Results clearly explain the existence of unstable intervals whose percentage is not negligible. Therefore, in this study I focus on unstable intervals in the above data to develop a prediction model for travel time over each interval.

4.5 Travel Time Variation over Unstable Intervals

This stage includes advanced analysis of the variability of travel time over each unstable interval for model development. This analysis consists of three statistical analyses to confirm the variations in travel time between of the eight time periods in a day, in the same time periods over days and the correlation of the travel time between adjacent time periods in a day.

4.5.1 Variability of Bus Travel Time between Time Periods

To evaluate the relationship between travel time over unstable intervals and traffic conditions over the intervals between adjacent bus stops of all routes, I carried out an exploratory analysis of the variability of travel time between time periods over days. In addition, this analysis confirms the assumption that travel time is strongly correlated with the time period in a day. During off-peak hour periods such as early morning (EM), midday (MD), and late night (LN) both traffic volumes and travel time show a normal distribution. In contrast, during peak-hour periods such as morning peak (MP), late morning (LM), and evening (E) the traffic volume increases dramatically and causes the travel time to increase. Therefore, the travel time of buses in

different time periods in a day should be different [53], [56]. This may help to capture the traffic conditions on that particular day and incorporate events such as accidents and route diversions.

Furthermore, a t-test was performed to analyze the travel time patterns on weekdays, with the statistical tests performed at the $\sigma = 0.05$ level of significance. The test compared the average of travel times for the eight time periods in a day and I serve the average of travel time for five weekdays to check whether the difference in the mean of the pairs of time periods is zero or not. The results of the comparison are shown in Table 4.1[1], including pooled standard deviation, f-values and p-values for the five weekdays for the inbound and the outbound directions.

TABLE 4.1: Periodical variance of travel time over unstable intervals for weekdays

Day	Inbound			Outbound		
	Pooled StDev (ln)	**F-Value**	**P-Value**	**Pooled StDev (ln)**	**F-Value**	**P-Value**
Mon	0.339	21.35	0.00	0.329	43.397	0.00
Tue	0.346	20.512	0.00	0.33	37.093	0.00
Wed	0.344	19.642	0.00	0.333	43.699	0.00
Thu	0.347	21.733	0.00	0.338	43.004	0.00
Fri	0.351	26.551	0.00	0.342	46.791	0.00

The results show that for the five weekdays p-values $<= 0.05$, indicating that there is a statistically significant difference in the average travel time among the eight time periods on the five weekdays. Further, on Thursday and Friday the pooled standard deviation values are higher than for other days for the inbound and the outbound directions. It is true that travel time over unstable intervals between adjacent bus stops may vary during the day on all routes. In the morning peak (MP), late morning (LM), and evening

[1]All the data has been transformed using natural logarithm to make data conform to normality distribution.

(E) on weekdays, bus travel time may significantly increase due to the heavy traffic volume, accidents or other events.

4.5.2 Variability of Travel Time over Unstable Intervals In Each Time Period over Past Several Days

Travel time variability in the same time period between days is caused by a number of variables whose impact cannot be anticipated by bus . The most common causes of travel time variability are temporal demand differences between off-peak and peak-hour [86]. The patterns of the variability can be typically classified into recurrent and non-recurrent variability. The recurrent variability is a result of insufficient capacity of roads, such as traffic congestion in peak-hour periods, while the non-recurrent variability is caused by transient events and sources of unexpected congestion, including accidents, inclement weather, construction and special events [53], [75].

Therefore, in the second part of the data analysis, I verified the daily variance of travel time over unstable intervals. In this work, analysis of variance (ANOVA) was further conducted to compare the average of travel time over unstable intervals in the same time period between different weekdays.

TABLE 4.2: Daily variance of travel time over unstable intervals

Time	Inbound		Outbound	
Period	**F-Value**	**P-Value**	**F-Value**	**P-Value**
EM	0.325	0.997	1.005	0.451
MP	1.380	**0.015**	6.615	**0.000**
LM	0.462	0.977	1.953	**0.008**
MD	0.810	0.697	1.034	0.416
EA	0.611	0.901	1.395	0.117
AP	1.483	**0.003**	3.010	**0.000**
E	2.577	**0.000**	4.435	**0.000**
LN	0.686	0.836	0.539	0.947

The results of the variability analysis of travel time obtained by using One-way ANOVA in SPSS software packageare shown in Table 4.2, where 0.000 denotes less than 0.05. The results show that some specific time periods, especially in peak-hour periods: morning peak (MP), afternoon peak (AP) and evening (E) for both the inbound and the outbound directions have a p-value smaller than 0.05, although travel time in the late morning (LM) for the outbound direction also has a p-value smaller than 0.05. Interestingly, in the other time periods there are p-values greater than 0.005, even though the target intervals are unstable ones.

The results indicate that historical average travel times in the off-peak periods seem to have recurrent properties and to help in prediction of travel time even over the unstable intervals. The results also support our assumptions that travel times over intervals in the same time period in peak-hour have non-recurrent properties and in off-peak-hour tends to have recurrent properties.

4.5.3 Correlation between Time Periods

As we know, a bus is not operated in the same way as other vehicles. Even though a bus delay may have been caused by a traffic jam or accidents, the bus cannot accelerate its speed to adjust to the delay because the bus has to follow the speed policy for the road and cannot change the route that has been determined by the time schedule [9], [49]. These facts suggest that the bus travel time in several time periods must have a correlation with that in the other time periods in a day [33], [39]. Because of this, I conducted experiments in this study to explore the correlations of travel time between the eight time periods in a day by using a statistical test (t-test).

To determine the effect of correlation in each time period, I calculate the Pearson Correlation Coefficients for all the combinations of paired different

time periods using the SPSS software. As the value is closer to +1 or -1, the two selected time periods are more closely related. Conversely, if the values are close to 0, the two selected time periods show weak or no correlation.

TABLE 4.3: Correlation between time periods of a day. In/Out denotes the inbound or the outbound directions

Time Period	Direction (In/Out)	EM	MP	LM	MD	EA	AP	E	LN
EM	In	1	**.78**	.73	.69	.76	.72	.68	.70
	Out	1							
MP	In		1	**.77**	.68	.71	.74	.72	.73
	Out	**.60**	1						
LM	In			1	**.82**	**.83**	.79	.76	.73
	Out	.59	**.79**	1					
MD	In				1	**.77**	.71	.67	.68
	Out	.49	.69	**.82**	1				
EA	In					1	**.84**	.76	.78
	Out	.52	.73	**.84**	**.81**	1			
AP	In						1	**.80**	.76
	Out	.54	.75	.79	.73	**.81**	1		
E	In							1	**.78**
	Out	.48	.74	.78	.71	**.82**	**.79**	1	
LN	In								1
	Out	.51	.71	**.80**	.73	**.79**	.75	**.78**	1

As shown in Table 4.3, the results mostly illustrate strong correlation between adjacent time periods. This means that travel time in the previous time periods is a very significant factor in determining the travel time in the later time periods. Because of that, I use them to predict travel time dynamically.

4.6 Factors Affecting for Prediction Travel Time Over Unstable Intervals

The empirical results of the variability exploration of the data include analysis of the average travel time over unstable intervals for the eight time periods in a day, as shown in Table 4.1. The results shown in Table 4.2 display

the variance of travel time in each time period over days. This table also includes the mean standard deviation of travel time in each time period of each segment (between time periods and between days).

In order to analyze how travel time over unstable intervals is correlated with the time period of a day, all correlations were taken as positive values, as shown in Table 4.3. There is a higher correlation between two time periods when the two time periods are near to each other .

Therefore, in fact there are two significant factors influencing the bus travel time over each interval between adjacent bus stops i.e. variations of travel time between time periods in a day, in the same time periods between days and correlation of travel time between time periods in a day. Thus, this study revealed that two types of travel time factors affect the prediction of travel time over unstable intervals and then uses the two type factors as input variables (independent variables) in the prediction model. The independent variables are defined in Section 5.3.2.

4.7 Summary

In summary, this analysis captured the variability of travel time over unstable intervals between time periods in a day, with the variability varied during the day. On the other hand, the average travel time within off-peak periods was relatively stable (variability of travel time tends to be recurrent), while the average travel time within peak-hour often varied every day (variability tends to be non-recurrent), and the correlation of travel time between adjacent time periods was relatively higher than that between non-adjacent time periods.

It's an important finding that travel time in the same time period may vary between days, indicating that we should use travel time in the time periods before the current one because the travel time in the previous time

periods is a very significant factor in determining the travel time in the later time periods and is thus a useful factor in predicting the travel time in those periods.

Chapter 5

Build a Prediction Model

5.1 Introduction

The algorithms used for travel time prediction in this study are based on dynamic traffic models. To begin with, a prediction model was developed to characterize whether the condition of bus travel time between time periods in a day over days shows recurrent or non-recurrent variability. There are three results I obtained from the statistical analysis of travel time over unstable intervals as mentioned in Section 4. First, the characteristics of travel time between adjacent bus stops may vary between time periods in a day. Second, the daily variance of travel time tends to be recurrent or non-recurrent. Third, there are strong correlations of travel time between time periods in a day. Based on the results of the analyses, I employed two types of input data: dynamic average travel time (DATT) which is travel time in the time period right before the current one and historical average travel time (HATT) in the same time period over the past several days to build our prediction model.

Next, I developed nonlinear dynamical models for predicting bus travel time over each unstable interval between adjacent bus stops in each of seven time periods in a day (omitting the EM period) . The proposed method basically utilizes time series methods based on Artificial Neural Network (ANN), Support Vector Machine Regression (SVR) and Random Forest (RF). I conducted experiments using bus probe data collected from November 21st to

December 20th, 2013.

5.2 Time Series Data Analysis

The averages of bus travel time over unstable intervals were obtained from all bus routes in each of eight time periods of a day for 20 days. The time series data runs from November 21 to December 20, 2013 (just for weekdays). In addition, in each day there are eight-time periods i.e., early morning (EM), morning peak (MP), late morning (LM), midday (MD), early afternoon (EA), afternoon peak (AP), evening (E) and late night (LN); the definition of the ranges of each time period is described in Table 3.2, Chapter 4.

Figures 5.1 and 5.2 shows the average of travel time over unstable intervals between adjacent bus stops. Each point in the graph is the mean of bus travel time in each of the eight time periods in a day over days. The vertical axes in the time series graphs represent the average of bus travel time between adjacent bus stops over unstable intervals. The horizontal axis represents the time period of each day.

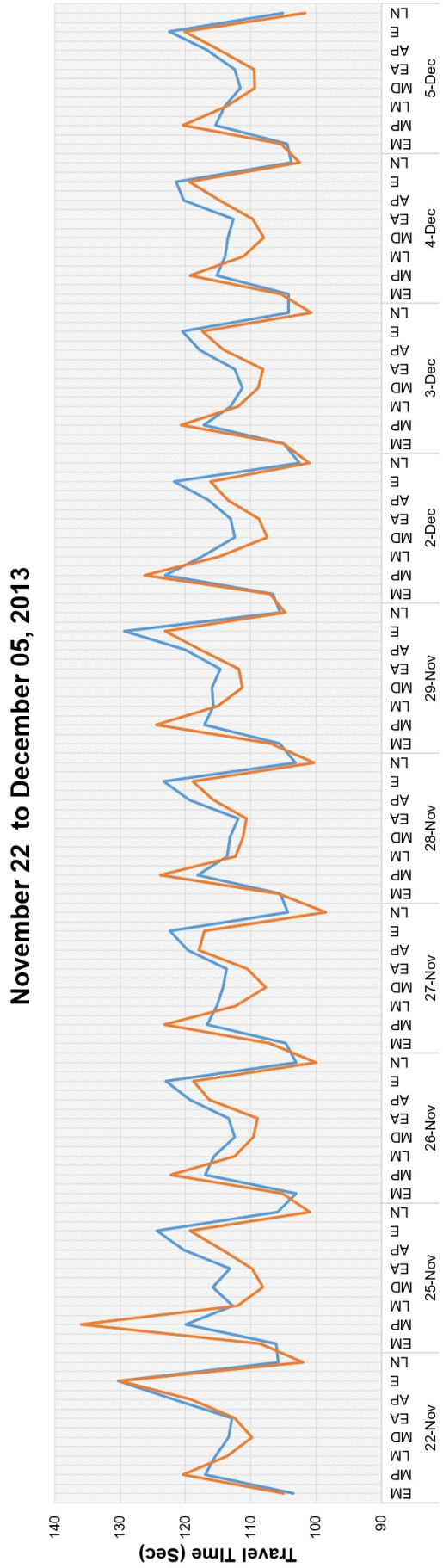

FIGURE 5.1: Observed average travel time over unstable intervals for weekdays

In Figure 5.1 and Figure 5.2 the time series provided some insight into the variability, trend and cyclic behavior of bus travel time. In general, the graph showed a greater increase in travel times in peak-hour periods than for other time periods in each day. However, off-peak periods tend to show the same values on different days (regular) .

The graphs showed that the mean travel time increased from the early morning (EM) towards the morning peak (MP) and decreased from evening (E) to late night (LN), which is very typical of observations of traffic conditions. The graphs showed higher travel time in the first week in the evening (E) period for the inbound direction and in the morning peak (MP) for the outbound direction, and the second week also showed higher travel time occurring in the evening (E) period for the inbound direction and in the morning peak (MP) periods for the outbound direction.

By contrast, in the third week the higher travel time occurred in the evening (E) period for the inbound direction and also in the evening (E) for the outbound direction. In the fifth week the higher travel time occurred in the evening (E) period for the inbound direction and in the morning peak (MP) period for the outbound direction, as for the first and second weeks. Similar time series trends in several time periods were evident in the early morning (EM), midday (MD), early afternoon (EA) and in the late night (LN) for the inbound and the outbound directions.

In addition, as described in Section 4.5, it was observed that the cyclic behavior is more prominent and the coefficient of variation is high-variance. The variability of travel time over unstable intervals comes from various sources, which can be categorized into two categories: regular variations (recurrent), e.g. variation between time periods day to day, and irregular condition variations (non-recurrent), e.g. when the same time periods show significant differences/random variations between days. So, with known regular and irregular conditions, dependent variations, and correlations of

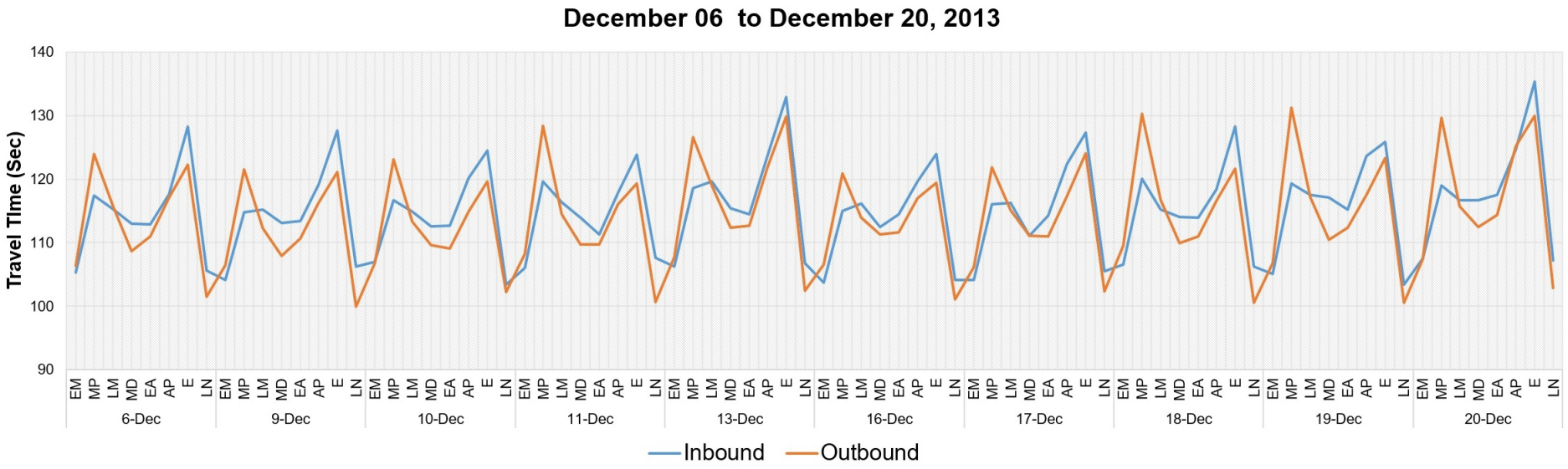

FIGURE 5.2: Observed average travel time over unstable intervals for weekdays

travel time between time periods in a day, it may be possible to predict the variability of travel time over unstable intervals of both categories.

5.3 Prediction Model Under Recurrent and Non-recurrent Variability

Although extensive research has been done on bus travel time variability, several challenges still need to be addressed. One of these challenges is how to address spatial and temporal correlations among the regular (recurrent) and irregular (non-recurrent) variations described in Section 5.2.

In this work, I carried out experiments to predict bus travel time over each unstable interval between adjacent bus stops using the SVR, ANN and RF methods. Then, for the input variables (independent variables) I used two types of input data: DATT and HATT, with the aim of capturing variability with the spatial and temporal correlations, especially for variability due to regular and irregular conditions. The input variables are described in Section 5.3.2.

5.3.1 Experimental Setup

In the experiments, prediction of travel time was performed in each of seven time periods in a day i.e., morning peak (MP), late morning (LM), midday (MD), early afternoon (EA), afternoon (AP), evening (E) and late night(LN) for the five weekdays. Because the input variable DATT uses the travel time observed in the time period just before the current one and there is no time period before early morning (EM), in this experiment I omitted the early morning (EM) period.

Dealing with ANN development, I used a radial basis function (RBF) architecture and a hidden layer feedforward network that is the most widely

used model for time series modeling and prediction. A feedforward neural network consists of three layers: one input layer, one hidden layer and one output layer. I chose the optimum number of neurons, which was 20 neurons, by trial and error, and then fed into the network the values for the past 20 days of the time series data.

A Nonlinear Autoregressive Network with eXogenous input (NARX) was chosen. The number of hidden neurons is 10, and the number of delays is 2. Using MATLAB software, I conducted MLP training with the Levenberg-Marquardt algorithm. I apply $Series - parallel(SP)mode$ to actual values of the target series data in order to form the regression of the target series data, and to minimize over-fitting in the training process after every n epoch, I perform validation using the validation data. To measure the model, I used Mean Squared Error (MSE) in training and testing. Using MSE in training the ANN model is a standard method, and MSE is a default indicator in training the network. The neural network model with the smallest MSE value is considered to be the best model. Starting from this performance criterion, the optimized NARX is obtained and used to perform the prediction of travel time in each time period of a day. I used a multilayer perceptron (MLP) with a single hidden layer to approximate any bounded continuous function. The MLP contains one or more layers of hidden units. The hidden units enable the MLP to learn complex tasks and meaningful features from the input/output relationships. Moreover, a high degree of connectivity between the MLP layers is determined by the weights of the network.

On the other hand, for SVR and RF, I used WEKA version 3.8. as described in Section 3.4. for SVR, RBF was selected as the kernel function in this study. Other parameters are C : 1.0, ε : 1.0E-12, and ε parameter tolerance: 0.001.

For RF observations with a tree default = 1.0; features per node scalar

default = nvars/3; maximum tree depth = unlimited, and the method of calculating variable importance = 0,1. In addition, most studies have noticed that increasing the number of trees does not decrease the predictive performance [67]. So, in this study, I used M = 500 trees, which is equal to their default values in the Weka RF package, and a node size = 5, while the number of features used for training at each node split is *mtry*. The parameter *mtry* is controlled during the validation phase to avoid over-fitting of data.

The optimal performance parameters for SVR, ANN and RF were found by calculating the average performance of the training models, while the search for the optimal value of the parameters was performed in a grid. Root Mean Squared Error (RMSE) is used to measure the performances. The optimal parameters for minimizing RMSE are selected. On the other hand, using SVR, ANN and RF for one-step ahead time series prediction is straightforward and similar to the methods that can be used for the time series approach.

5.3.2 Input variables and Training Data

In this section, I discuss input variables of the model. The inputs consider travel time in each time period. As we know, bus travel time will vary according to time period. Especially in the morning peak (MP) and afternoon peak (AP) periods, bus travel time will increase significantly, as described in Section 4.6.

The analysis established that there are two significant factors that influence bus travel time over each unstable interval between adjacent bus stops, namely variations in travel time in the same time periods over days and correlations of travel time between time periods in a day. Therefore, I define the following two types of travel time, which are expressed as input variables (independent variables) for the training data.

1. **DATT** is expected to adjust the prediction of travel time in the current time period using the travel time observed in the period just before the current one. For instance, I predict travel time in the morning peak (MP) using the travel time observed in the early morning (EM). It makes an additional contribution to detecting unexpected dynamic events than is possible only using HATT.

2. **HATT** denotes the average travel time in the same time period during the past several days. It is an important input variable of the prediction model because travel time over intervals in some time periods tends to have (regular) properties that recur on other days.

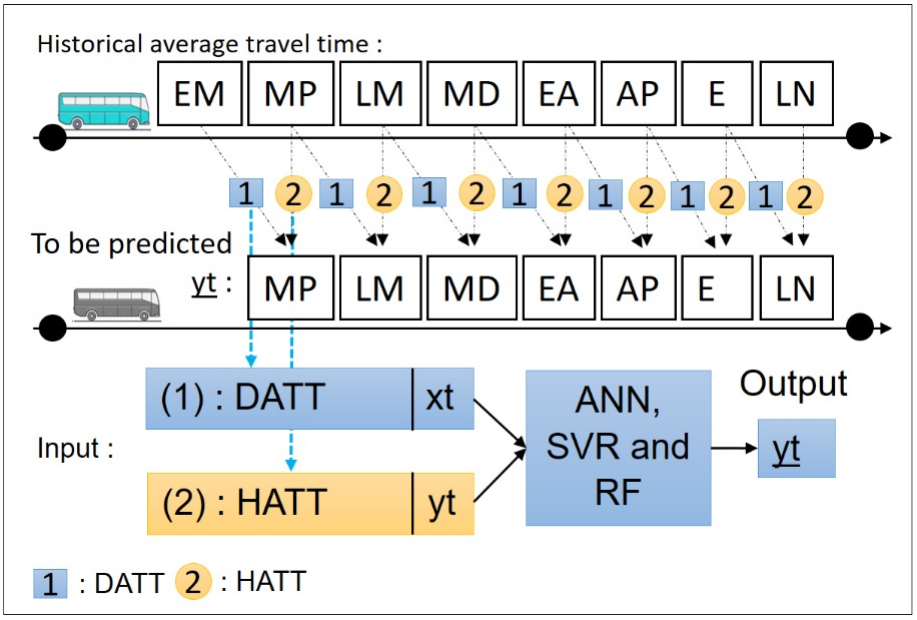

FIGURE 5.3: Input variables of the model

Figure 5.3 illustrates how the model uses input variables to predict travel time. Arrows with digit 1 in a box and circled digit 2 denote an input of DATT for the time period just before the current one, and an input of HATT for the current time period during past several days, respectively.

Next, in the training data, I used both types of input data, DATT and HATT, in the prediction model. Prediction is accomplished in five iterations,

with each iteration producing one step ahead prediction results (for the seven time periods of travel in a day). For each model: SVR, ANN and RF, in the first prediction step, I used 14 days of data as training data, 1 day of data as validation data, and the next available day from the 5 days of testing data. In the training process, predicted results were evaluated by RMSE conducting out-of-sample prediction with validation data, and when the RMSE value fell below the threshold, the training was stopped to avoid over-training.

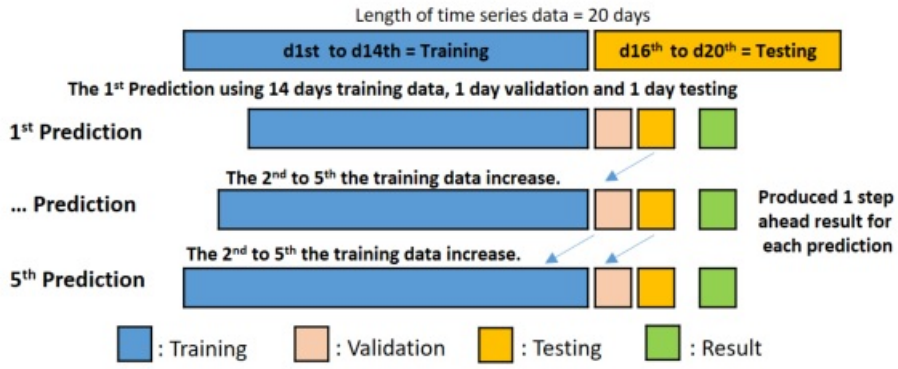

FIGURE 5.4: Training models

I repeated all the above steps five times selecting another day's data from the 5-day test data. In the next iteration step, I added the validation data in the previous iteration step into the training data and used the test data in the previous iteration step as the validation data. Thus, the amount of training data increased by one day in each iteration step. The iterations of the prediction model are shown in Figure 5.4.

5.3.3 Performance of Prediction Models

When building prediction models, the primary goal is to make a model that most accurately predicts the desired target value for actual data. To measure the model error, I used mean absolute percentage error (MAPE) as described in Section 3.4.5. I focused on seven time periods, excluding the first of the

eight time periods in a day, early morning (EM).. This is because DATT relies on the travel time observed in the time period just before the current one, and there is no time period before early morning (EM). I just used HATT for EM.

Figure 5.5 presents the MAPE values of the prediction models in each time period for five days in performing one-step-ahead prediction for weekdays. First, I observed the prediction results for five days. For the inbound direction on Monday in morning peak (MP), the RF model obtained the lowest prediction error of approximately 6.72% and on Wednesday in morning peak (MP) the SVR model obtained the lowest prediction error of approximately 6.70%. Next on Tuesday, Thursday and Friday always in early afternoon (EA) the ANN model obtained the lowest prediction error of approximately 6.94%, 6.57% and 6.31% respectively. For the outbound direction, the RF model obtained the lowest prediction error on Monday for midday (MD) of approximately 5.46%, the SVR model was next, obtaining the lowest prediction error for several days i.e, Tuesday in morning peak (MP), Wednesday in afternoon peak (AP), Thursday in morning peak (MP) and Friday in late morning (LM), with values of approximately 5.55%, 7.37%, 5.02% and 6.69%, respectively.

In general, the observations indicate that there is no significant difference in the distribution of MAPE among the three models. However, it can be clearly seen for the inbound direction that the ANN model outperformed the SVR and RF models for the prediction of travel time on several days, especially in time periods with recurrent variability i.e., early afternoon (EA). On the other hand, for the inbound and the outbound directions, the SVR and RF models give better prediction results compared to the ANN model in predicting travel time in time periods with non-recurrent variability i.e., in morning peak (MP), afternoon peak (AP) and evening (E). Recurrent and non-recurrent variability of travel time were observed as described in Section 4.5.2.

Second, in order to clarify the above discussion, I carried out further analysis on overall average MAPE for each time period for the three models. The analysis results can be seen in Table 5.1.

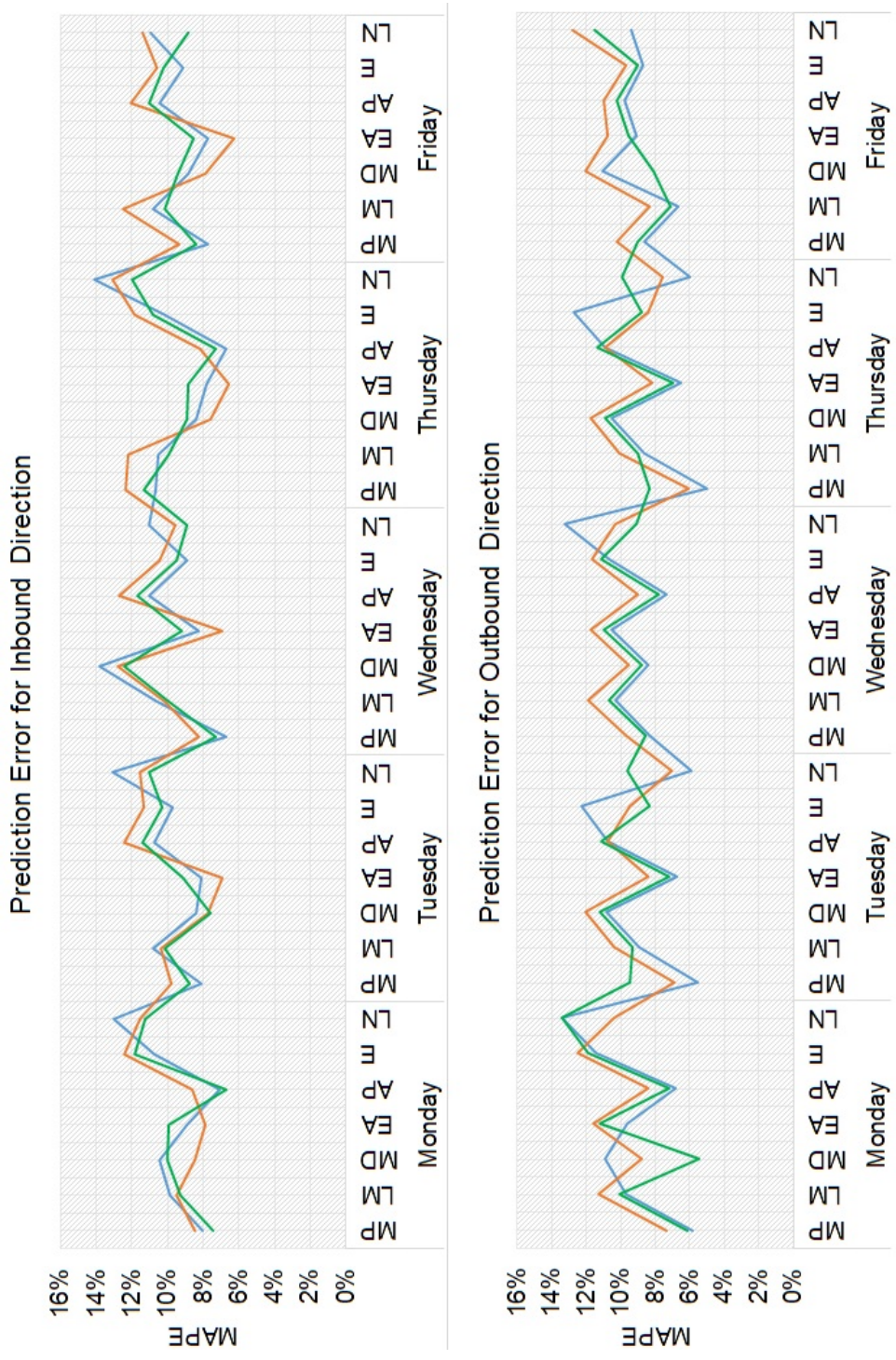

FIGURE 5.5: Prediction error for the inbound and outbound directions

TABLE 5.1: Average MAPE of prediction error

Models	Direct.	MP	LM	MD	EA	AP	E	LN
SVR	In	8.3%	10.5%	9.9%	8.2%	9.2%	9.7%	12.5%
	Out	6.7%	8.9%	10.3%	8.5%	9.1%	11.2%	9.6%
ANN	In	9.7%	10.9%	8.9%	6.93%	10.8%	11.3%	11.4%
	Out	8.0%	10.4%	10.8%	10.16%	10%	10.4%	9.6%
RF	In	8.6%	9.9%	9.7%	9.1%	9.6%	10.5%	10.4%
	Out	8.3%	9.2%	8.9%	9.2%	9.6%	9.8%	10.7%

As can be seen in Table 5.1, the SVM and the RF models have the lowest MAPE value in periods of non-recurrent variability i.e., morning peak (MP), afternoon peak (AP) and evening (E). However, the SVR model obtained worse prediction results in periods of recurrent variability for the inbound direction i.e., late night (LN) and midday (MD), with values of approximately 9.99% and 12.45% respectively. The RF model also obtained worse prediction results in late night (LN) for the outbound direction, compared with the ANN model. In summary, the SVR and RF models outperform the ANN model when travel time shows non-recurrent variability from day to day, but the results of all three models showed acceptable performance and are in the reasonable error range in predicting the travel time over unstable intervals, especially for recurrent and non-recurrent variability .

5.3.4 Assessing the Significance

Next, it was necessary to conduct a second test to examine the significant differences between the three models, especially the ability of DATT and HATT data to capture recurrent variability such as traffic congestion in peak-hour periods and non-recurrent variability caused by transient events, including accidents, inclement weather, construction and special events.

I conducted analysis variance (ANOVA) using the SPSS Software package to compare the error of prediction results for all intervals, with the aim of

seeing if there are any significant differences among the three models in each time period for all days. In addition, I used interpretation with a significance level of 0.05% to increase confidence in the difference.

TABLE 5.2: ANOVA t-test of prediction error between SVR, ANN and RF

Comparison	Inbound		Outbound	
Results	**T-value**	**P-value**	**T-value**	**P-value**
MP	1.014	.392	1.501	.262
LM	2.057	.171	1.744	.216
MD	.352	.710	1.669	.229
EA	21.621	**.000**	.996	.398
AP	.678	.526	.331	.724
E	4.753	**.030**	.895	.434
LN	2.781	.102	.278	.762

The results of the ANOVA test are shown in Table 5.2. For the inbound direction, there are only two time period p-values that are less than 0.05, which indicates that there are significant differences in results for those two time period, namely the early afternoon (EA) and (E) evening periods. For the outbound direction, however, there are no p-values less than 0.05, which means that there were no differences in prediction results for any time periods in that direction.

Therefore, considering the above results, I can say that there are no significant differences for the outbound direction among the three models in predicting unstable intervals that include recurrent and non-recurrent variability of travel time. However, there are two-time periods for the inbound direction which show significant differences. I assume that these differences occur because of the unpredictable traffic flow in congested conditions, especially in peak-hour periods, and I investigate this further in Section 6.

5.4 Summary

It is a challenge to predict bus travel time over unstable intervals between adjacent bus stops which include significant variability (recurrent or non-recurrent). Using machine learning methods i.e., SVR, ANN and RF I built prediction models without using weather data, traffic data etc. Because of that I distinguished stable and unstable intervals and carried out an exploratory data analysis on the variability of unstable intervals that indirectly represents the characteristics of traffic conditions (travel time).

Further, based on the exploratory data analysis, I determined to use two types of input data as independent variables, namely HATT (historical average travel time data for the same time periods) and DATT (dynamic average travel time data in the time period just before the current one) and the prediction model was built based on the time series method approach.

Experimental results showed that the SVR and RF models outperform the ANN model in most cases, but there were no significant differences between the three models. Also, the results of the three models showed acceptable performance and are in the reasonable error range in predicting the travel time over unstable intervals. The results also indicated that the three models accurately and dynamically predicted travel time over unstable intervals in each time period in a day for 5-days. This means that bus travel time can be reasonably estimated using both DATT data and HATT data for unstable intervals with recurrent or non-recurrent variability.

Chapter 6

Prediction Models Based on Off-peak and Peak Hours

6.1 Introduction

This chapter aims to build a framework for predicting bus travel time based on the off-peak and peak-hour periods in a day. Moreover, the goal of this chapter is to develop a model that predicts travel time in terms of traffic parameters and thereby evaluates the impact of changes in the input variables.

The literature review indicated that predicting travel time using off-peak and peak-hour periods in a day as parameters is more challenging than other methods. In other words, an input variable is important in predicting travel time [85] for unstable intervals because the performance of the prediction model often deteriorates when the number of input variables (independent variables) increases, i.e when the dimension of the input space increases. This has been referred to in the literature as the problem of dimensionality [56]. This phenomenon is due to the selection of irrelevant input variables for modeling, which may increase model complexity and hence lead to poor generalization. Increasing the number of input variables also leads to the necessity for more parameter training. Thus, it becomes necessary to understand the input and output relationships between the independent variables.

Therefore, in this model the input variables are manipulated before choosing a prediction algorithm, and then I differentiate travel time in each of the eight time periods in a day into off-peak and peak-hour periods. Next, I use two types of machine learning techniques, ANN and SVR, to demonstrate the impact of types of input variables in predicting the travel time over each unstable interval.

6.2 Establish Prediction Model

It was observed in Section 4.5.2 that the analysis captured the fact that travel time variability varied between days, as shown in Table 4.2. The results show that for the inbound direction there are three time periods whose p-values are less than 0.05 and present significant differences between days, namely morning peak (MP), afternoon peak (AP) and evening (E). For the outbound direction, there are four time periods whose p-values are less than 0.05 and present significant differences between days, namely morning peak (MP), late morning (LM), afternoon peak (AP) and evening (E). This means that the significant differences in travel time occur during peak-hour periods. On the other hand, during off-peak periods, the travel time in the same time periods is fairly constant over the weekdays.

These results clearly indicate that the travel time variability between peak-hour periods and off-peak periods over days is significantly different. This is an important finding that suggests that independent variables determining the characteristics of the travel time variability should be considered. So it is necessary to establish a method that uses the input variables (HATT and DATT) selectively with a distinction between off-peak and peak-hour periods. Next, another point deserving attention is that there are strong or moderate correlations of travel time between adjacent time periods as described

in Section 4.5.3. This is a fact there is variability closely related with regular and irregular of the travel time over each unstable interval over days.

Thus, in this work, I divide the eight time periods in a day into two categories: off-peak periods, namely early morning (EM), midday (MD), early afternoon (EA), and late night (LN), and peak-hour periods i.e., morning peak (MP), late morning (LM), afternoon peak (AP), and evening (E). Then I build prediction models over each unstable interval using the following scheme:

1. When predicting travel time over each unstable interval in the off-peak periods, i.e. early morning (EM), midday (MD), early afternoon (EA) and late night (LN), only one input variable, HATT, is used.

2. Otherwise, in the peak-hour periods, i.e. morning peak (MP), late morning (LM), afternoon peak (AP) and evening (E), two types of input variables, DATT and HATT, are used.

DATT and HATT refer to average travel time in the period right before the current one and average travel time in the same time period over the past several days, repectively. DATT uses strong correlations of travel time between adjacent time periods and HATT uses those of travel time in the same time periods over days, especially weekdays. Here, for the off-peak periods, I only use HATT as shown in Figure 6.2. This is different from the method mentioned in Section 5.3, which uniformly uses two types of input data (input variables) for both the peak-hour and off-peak periods.

Furthermore, during preprocessing training, the walk-forward testing routine is used to divide each data set into five overlapping ones as shown in Figure 6.1. I used data for a total of 20 weekdays, numbering each day from 1 to 20. In the first iteration, I used 14 days of data from the 1st day to the 14th day data as training data, the next 1 day of data as validation data, and the following 1 day of data as testing data. Then, in the next iteration, I shift 1 day forward in the data series without changing the size of the training,

validation or testing data and repeat until the 5th iteration. In each iteration, I calculated MAE and RMSE and took the average of the MAE and RMSE results obtained in five iterations.

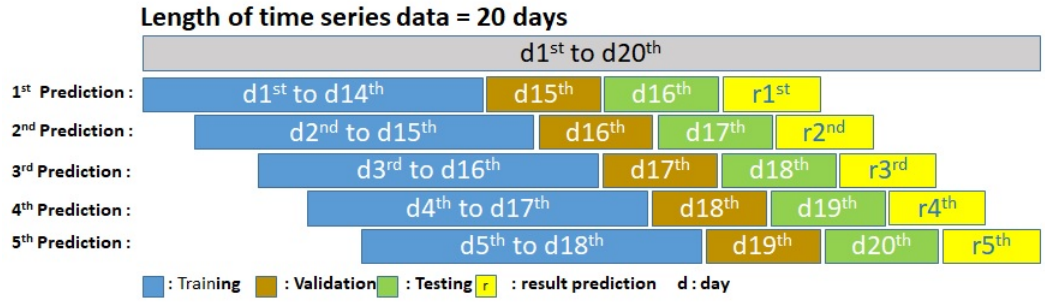

FIGURE 6.1: Establishing the training model

6.3 Experimental Setup

Next, two types of methods, Artificial Neural Network (ANN) and Support Vector Regression (SVR), are used to estimate travel time in each of eight time periods in a dayfrom the above data. I built the ANN model using the MATLAB software and the SVR model using libsvm-3.22 provided by Chih-Chung Chang and Chih-Jen Lin [15].

In addition, when developing the ANN model, I used a radial basis function (RBF) kernel and a feedforward network, which is the most widely used model for time series modeling and prediction. The feedforward neural network consists of three layers: one input layer, one hidden layer and one output layer. I chose the optimum number of neurons by trial and error, which was 20 neurons; I fed into the network the values for the past 20 days of the time series data. Next, I also developed the SVR model with an RBF kernel function.

For measurement of performance C in the training process, I select the parameters of the RBF network ($\lambda\epsilon$) taking the minimum error for the SVR model [79]. Likewise, I evaluated the parameters of the ANN and the SVR

models with emphasis on Root Mean Squared Error (RMSE) using the validation data set and selected the model parameters with the smallest RMSE values. Next, I selected Mean Absolute Percentage Error (MAPE) to measure the prediction accuracy of the two models. RMSE and MAPE are defined in Section 3.4.5.

Figure 6.2 shows the scheme of the two types of input data (independent variables) for peak-hour and off-peak periods in a day.

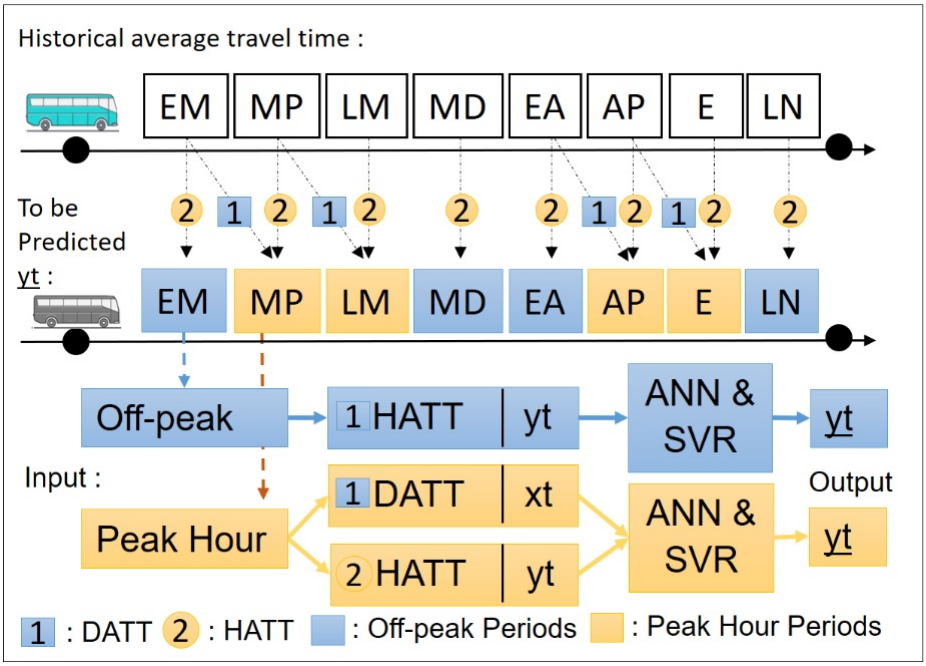

FIGURE 6.2: Scheme of the input data

6.4 Evaluation of Predictive Models

Figure 6.3 shows the results achieved in predicting travel time over unstable intervals for weekdays in off-peak and peak-hour periods in a day. It can be observed that by using two input variables in peak- hour periods and one input variable for off-peak periods, the prediction accuracy can effectively be improved. The results show that the ANN and SVR models in off-peak periods achieved a prediction error of less than 9% for the inbound and the

outbound direction, while, for the peak-hour periods, they achieved a prediction error of less than 8%.

However, in our experiment for the inbound direction, the ANN model obtained higher MAPE values for some days in off-peak periods, especially for midday (MD), i.e. Monday 8.94%, Tuesday 8.78%, Wednesday 8.97%, Thursday 8.40% and Friday 8.62%. At the same time, for the outbound direction, the ANN model also obtained higher MAPE values in off-peak periods, i.e. Monday 7.79%, Tuesday 8.03%, Wednesday 7.91%, Thursday 7.88% and Friday 8.00%.

Next, to evaluate the prediction performance, I conducted two experiments. First, I compared the prediction accuracy of the ANN and SVR models to find which model shows superior prediction performance. Tables 6.1 and 6.2 show the comparison results using average MAPE for all time periods (off-peak and peak-hour periods).

TABLE 6.1: Comparison for the off-peak periods

Time	**Inbound**		**Outbound**	
Period	ANN	SVM	ANN	SVM
EM	6.45%	6.49%	6.42%	6.40%
MD	8.74%	6.83%	7.92%	6.68%
EA	6.59%	6.45%	6.54%	6.29%
LN	6.84%	6.54%	7.12%	6.51%

TABLE 6.2: Comparison for the peak-hour periods

Time	**Inbound**		**Outbound**	
Period	ANN	SVM	ANN	SVM
MP	6.31%	6.40%	6.16%	6.09%
LM	6.62%	6.51%	6.63%	6.35%
AP	7.43%	6.40%	6.80%	6.41%
E	6.78%	6.31%	6.87%	6.48%

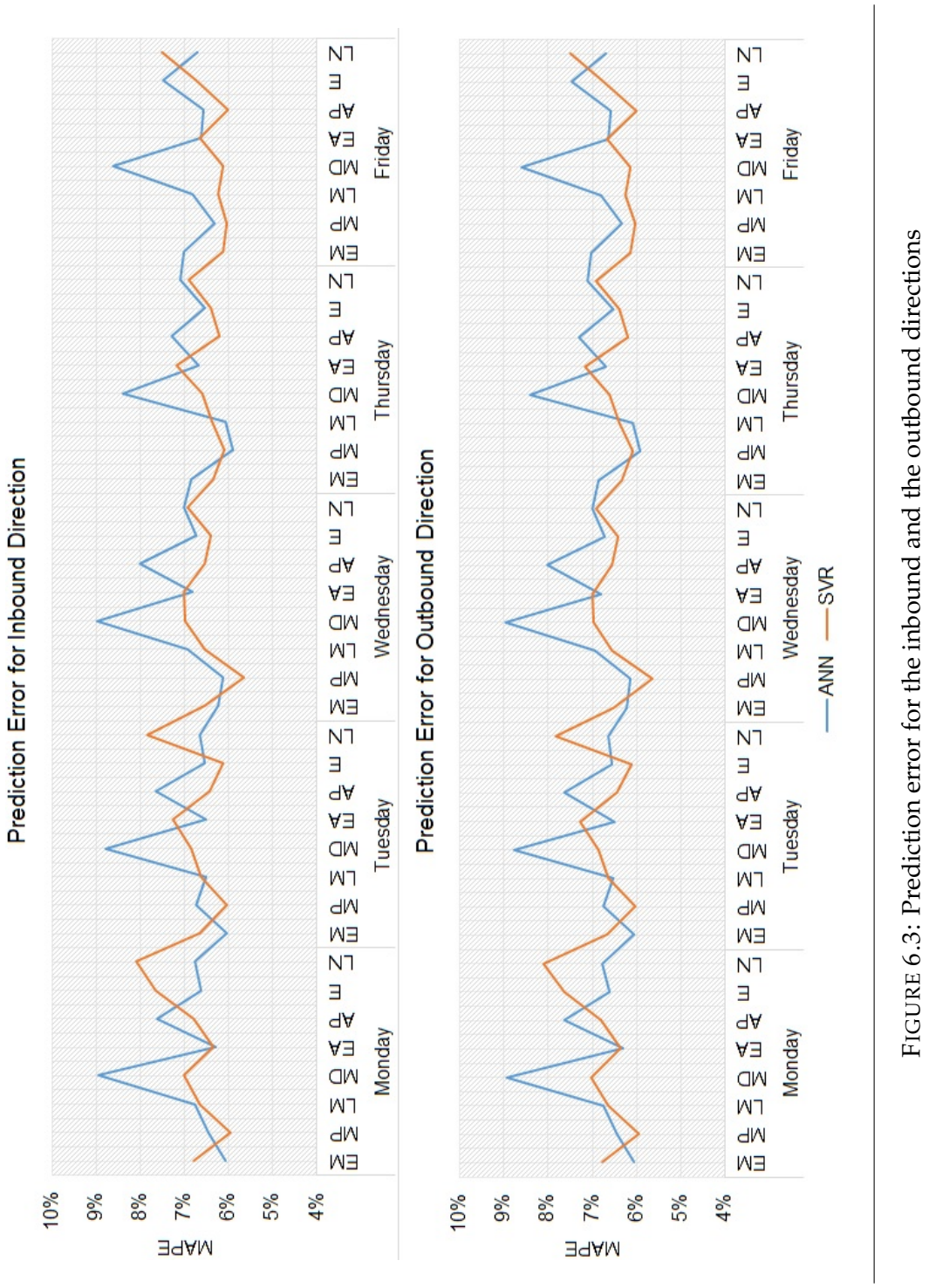

FIGURE 6.3: Prediction error for the inbound and the outbound directions

Second, I carried out two analyses to assess the sensitivity of the ANN and SVR models to the input variables and to examine the robustness of the two models. In the first analysis, our target is off-peak periods. Using only one type of input data, the SVR model had 7.46% of the average MAPE value in the late night (LN) for the inbound direction and 7.42% for the outbound direction, and also showed 7.31% in the Midday (MD) for the outbound direction. The ANN model, on the other hand, obtained 8.47% in the midday (MD) for the inbound direction and 7.92% for the outbound direction. Moreover, the ANN model also obtained 7.12% in the late night (LN) period, which is poor prediction performance. In the second analysis, using two input variables with the peak-hour periods as the target, the SVR model only obtained 6.66 % in the evening (E) for the inbound direction and 6.74% in the late morning (LM) for the outbound direction, while the ANN model obtained 7.43% in the afternoon peak (AP) period for the inbound direction and 6.87% in the evening (E) period for the outbound direction.

6.4.1 Assessing the Significance

In order to establish whether the prediction error differs between the ANN and SVR models, a paired sample t-test can be performed. Statistical significance is determined by looking at the p-value. The p-value gives the probability of observing the test results under the null hypothesis. However, in this comparison the cutoff value for determining statistical significance is a value of 0.05 or less.

The results of the comparison are shown in Tables 6.3 and 6.4. The tables include the results of the comparison of the Mean Absolute Percentage Error (MAPE) between ANN and SVR over unstable intervals in each time period for 5 days of prediction. First, focusing on the off-peak periods for the inbound direction, only for one time period is the p-value less than 0.05,

TABLE 6.3: Paired samples test for the off-peak periods

SVR	Inbound		Outbound	
& ANN	T-value	P-value	T-value	P-value
Pair 1 EM	-.139	.896	.079	.941
Pair 2 MD	16.777	.000	4.548	.010
Pair 3 EA	-2.155	.097	10.714	.000
Pair 4 LN	-1.943	.124	-2.230	.090

TABLE 6.4: Paired samples test for the peak-hour periods

SVR	Inbound		Outbound	
& ANN	T-value	P-value	T-value	P-value
Pair 1 MP	2.416	.073	.426	.692
Pair 2 LM	.807	.465	-.449	.677
Pair 3 AP	6.664	.003	2.923	.043
Pair 4 E	.388	.718	1.366	.244

namely the midday (M) period. On the other hand, for the outbound direction there are two time periods whose p-values are less than 0.05, namely the midday (M) and early afternoon (EA) periods. Next, in the peak hours, for the inbound and the outbound directions there is only one-time period, namely afternoon peak (AP) whose p-value is less than 0.05. As all results indicate, there is no significant difference between the two models.

6.4.2 Summary

I discussed the two prediction models for travel times over each unstable interval between adjacent bus stops considering eight time periods in a day. I built the two models (ANN and SVR) using real bus probe data. Before building the models, I conducted an exploratory analysis of the variability of travel time over each interval, and classified all intervals into stable and unstable ones. Next, focusing on the unstable intervals, I confirmed the variability of the travel time over each interval among the eight time periods in

a day, the variability in the same time period between weekdays, and the correlation of travel time between adjacent time periods in a day.

From the results, I proposed a method which uses two input variables selectively with a distinction between peak-hour periods and off-peak periods, considering the traffic characteristics over unstable intervals. Then, I evaluated the prediction performance of the two models in off-peak and peak-hour periods. Both of the two models showed an acceptable prediction performance for both types of periods. Although the SVR model had better prediction results than the ANN model in most of the time periods, there was no significant difference between the two models.

Chapter 7

Assessing the Performance of the Prediction Models

7.1 Introduction

In this chapter, I evaluate the performance of the models by conducting several comparison experiments. First, I conducted a comparison experiment between our proposed model and the model in a previous study [74]. Second, I compare the proposed method with the model in our previous study [5]. The aims of this section are to assess the performance of and significant differences between prediction models without attempting to identify a "true" or "best" model.

7.2 Comparison between the Proposed Model and Another Model

7.2.1 Experiment Setup

The performance of prediction models can be assessed using a variety of different methods and metrics. An alternative approach to identifying the relevance of the input variables and to exploring the impact of individual input variables on the model is to use comparison techniques. Therefore, to assess

the prediction performance of the model, I compared our model with a model proposed by Swardo et al. [74], which was developed as an ARIMA model. In addition, to predict travel time they only used average travel time data obtained from the preceding several days. I call their model a HATT-based Model.

I conducted experiments to compare our model with the HATT-based Model on their travel time prediction performance over each unstable interval without considering other factors, such as weather, road and traffic conditions. The prediction is performed for seven time periods in a day: morning peak (MP), late morning (LM), midday (MD), early afternoon (EA), afternoon peak (AP), evening (E) and late night (LN).

When conducting these model experiments, I used time series data covering 20 days and in which there are eight-time periods in each day to predict each of the eight time periods in a day listed above . Also, for both models, I conducted the experiment using the ANN method from the MATLAB software, specifically choosing the nonlinear autoregressive network with exogenous input (NARX), which is known as NARX recurrent neural network. When building both models, I used a multilayer perceptron (MLP) with a single hidden layer to approximate any bounded continuous functions. The hidden units enable the MLP to learn complex tasks and meaningful features from the input/output relationship, and I conducted the MLP training using the Levenberg-Marquardt algorithm. To measure training data for both the models, I used Mean Squared Error (MSE) in training and testing. The neural network model with the smallest MSE value is considered to be the best model. The MSE is defined as follows:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \ddot{y}_i)^2}{n}, \tag{7.1}$$

where y_i and $\ddot{y}_i$ are the ith observed and the predicted values, and n is the

number of observations. Starting from this performance criterion, the optimized NARX is obtained and used to perform the prediction of travel time in each time period of the day for both models. The experimental process for both models is as follows:

1. **Preparation of training data.**

   In the training phase for the model proposed here, I use two types of input data (input variables): DATT and HATT, where DATT represents the past exogenous values, x(t), and HATT represents the past values, y(t). The input variables of the proposed model are described in Figure 5.4 in Section 5.3.2. By contrast, for the HATT-based model, I just use one input variable, i.e HATT, which represents both x(t) and y(t) as shown in Figure 7.1.

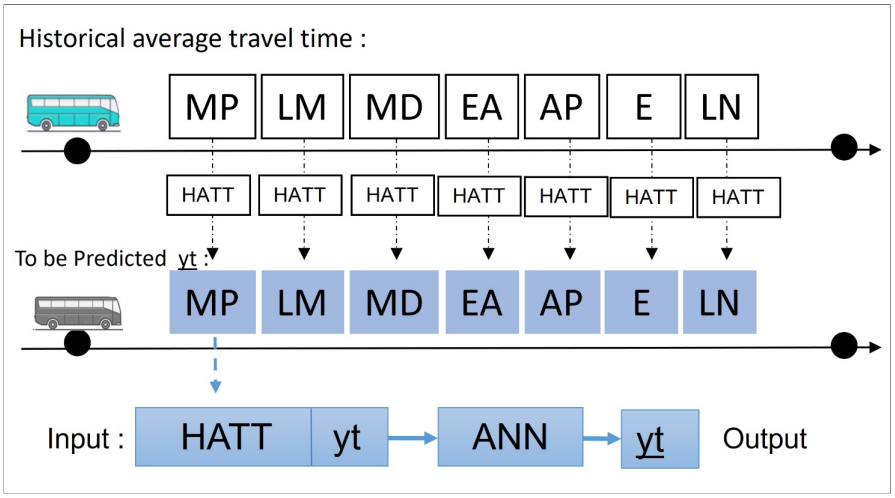

FIGURE 7.1: Variable input for the previous model

Next, both models in the first prediction step use 14 days of time series data for training, 1 day of data for validation and 1 day from 5 days of data for final testing. During the training data phase between the second and the fifth iteration of prediction, I add to the training data the validation data used in the previous iteration, and use the test data in the previous iteration process as the validation data in the current

iteration process as shown in Figure 6.1. In this way, the length of the training data increases by one day in each iteration. I use the MATLAB software in the neural network toolbox to process them. Figure 5.4 shows how to process training data in each iteration of the prediction.

2. **Configuring the network architecture.**

   I select the Narxnet network architecture (*narxnet*) defaults: the number of hidden neurons (*hiddenSizes*) is 10, row vector of increasing 0 or positive delays ($inputDelays$ and $feedbackDelays$) = 1:2, and training function ($trainFcn$) = $'trainlm'$. The aim is for the networks to learn to predict one step ahead in the time series of travel times, given the past values DATT and HATT.

3. **Training and validation.**

   In the training network of both models, I use the Levenberg-Marquardt algorithm in MATLAB trainlm. To minimize over-fitting in the training process after every n epoch, I perform validation using the validation data, calculate MSE as defined in equation 7.1, and select a small value of the MSE. This is not to adjust the weights of the network model, but just to verify whether there is any increase in accuracy during the training. Next, I select the network which takes the minimum error using the evaluation set to make predictions on the test set for the next step.

4. **Conversion of the NARX architecture.**

   The series-parallel NARX architecture is converted from an open to closed configuration by removing delays. Training and validation in step 3 are repeated by the SP NARX architecture. If the difference in MSE between the output and the target (HATT) becomes smaller than

a threshold, the process will go to step 5, but if the MSE takes on a large value, the process will go back to step 3 and repeat the training algorithm.

5. **Testing.**

   Because the relationship between past and future values of the network architecture does not change (*narxnet* default) [25], the testing model is designed directly from data. Thus, the system performance is tested using 5 days of data which have not been used in the training or validating phases.

6. **Final phase (Prediction).**

   I calculate the Mean Absolute Percentage Error (MAPE) of each prediction result by comparing the testing data. MAPE is defined in equation (3.19).

7.2.2 Measuring the Performance of Both Models

Performance measurement is generally defined as regular measurement of results, which generates reliable data on the effectiveness of the prediction model [72]. Therefore, using Mean Absolute Percentage Error (MAPE) as shown in Figure 7.2, I evaluate prediction performance. This comparison is an indicator that helps us to measure change in the prediction results over time.

From the graph as shown in Figure 7.2, for observed travel times with non-recurrent variability in the time periods, namely morning peak (MP), late morning (LM), afternoon peak (AP) and evening (E) for the inbound and the outbound directions, our prediction model can achieve less than 10 % error. In other words, our model using DATT, which denotes average travel time in the time period right before the current one, indirectly succeeded in

capturing unexpected events, compared with the HATT-based model, which only captured the historical average of travel time over the past several days.

However, the HATT-based model had better prediction results for two time periods, namely midday (MD) and late night (LN). As one might expect, in these periods, bus travel time is normally stable due to the traffic conditions. Overall, the results indicate that our model mostly obtained lower MAPE values than the HATT-based model, in particular for in peak-hour periods.

7.2.3 Measuring the Significance

Next, it was necessary to conduct another experiment to examine the effective characteristics of DATT data, such as to capture the non-linear variability of unstable intervals [4]. I compare the prediction results for all intervals. I conducted paired sample t-tests from the SPP Software package to see if there were any significant differences in each time period for all days between our model (proposed) and the HATT-based model.

As shown in Table 7.1, the results obtained for the inbound and outbound directions show that there are significant differences between our model and the HATT-based model.

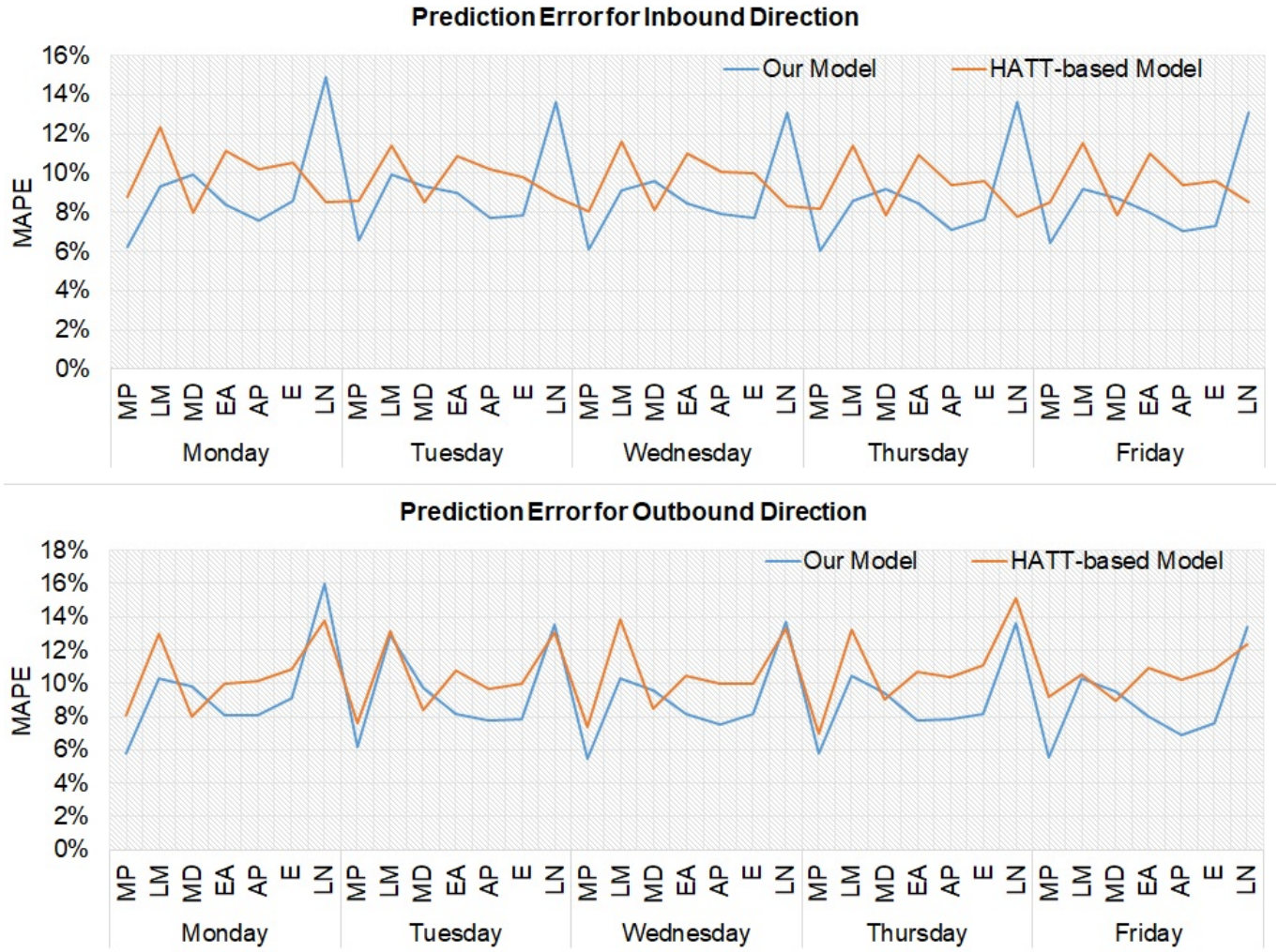

FIGURE 7.2: Comparison of prediction performance between proposed and HATT-based Models

TABLE 7.1: Paired sample tests between proposed and HATT-based models

Comparison	Inbound		Outbound	
	T-value	**P-value**	**T-value**	**P-value**
Pair 1 MP	-23.079	**.000**	-4.912	**.008**
Pair 2 LM	-9.255	**.001**	-2.782	.050
Pair 3 MD	6.152	**.004**	3.937	**.017**
Pair 4 EA	-13.230	**.000**	-12.180	**.000**
Pair 5 AP	-29.936	**.000**	-10.207	**.001**
Pair 6 E	-28.476	**.000**	-7.774	**.001**
Pair 7 LN	15.043	**.000**	.917	.411

For the inbound direction in each time period for all days, the p-value is less than 0.05, which means that there are significant differences between the models. Furthermore, for the outbound direction, there are two time periods, namely late morning (LM) on Tuesday and late night (LN) on Wednesday, when the p-values are greater than 0.05. For other outbound time periods, the p-values are less than 0.05.

In general, these observations agree with results from other studies indicating that ANN models have the ability to capture the complex non-linear relationship between travel time and the independent variables [35].

7.3 Comparison between the Proposed Model and that in Our Previous Study

7.3.1 Experiment Setup

In order to discover the characteristics of the training data in our previous study [5], I conducted further experiments predicting travel time for each time period in both the inbound and outbound directions according to the

training model in our previous study. Since the SVR model had the best performance in this study, I repeated the experiments five times as I conducted the experiments in Section 5.3.2.

To make a fair comparison between those earlier results and those from the current study, I conducted experiments on our previous model using time-series data for 20 days; each day is divided into eight time periods and I made predictions for seven of those time periods. Next, I conducted experiments using WEKA version 3.8 as described in Section 5.3. RBF was selected as the kernel function, and the parameters were C : 1.0, ε : 1.0E-12, and ε parameter tolerance: 0.001. In the training phase, I also used two sets of input data: DATT and HATT, for each of the two models, as described in Section 5.3.2.

In the first prediction step, I used 14 days of data as training data, 1 day of data as validation data, and 1 day of data as testing data. I prepared 5 days of data as testing data and selected 1 day's data from among them in each iteration without duplication. In the training process, predicted results were evaluated by RMSE, conducting out-of-sample prediction with validation data; when the RMSE value fell below the threshold, the training was stopped to avoid over-training. I repeated all the above steps five times, selecting another 1 day of data from the 5 days of test data.

In the next iteration step, I added the validation data in the previous iteration step into the training data and used the test data in the previous iteration step as the validation data. Thus, the quantity of training data increased by one day's worth of data in each iteration step. The iteration of the prediction model is shown in Figure 7.3. Unlike our previous study [5], this procedure does not add in the prediction result of the current step, but adds the validation data in the current step into the training data in the next step.

Further, I investigated if just adding the validation data used in the previous step into training data in the current step has any effect. To this end,

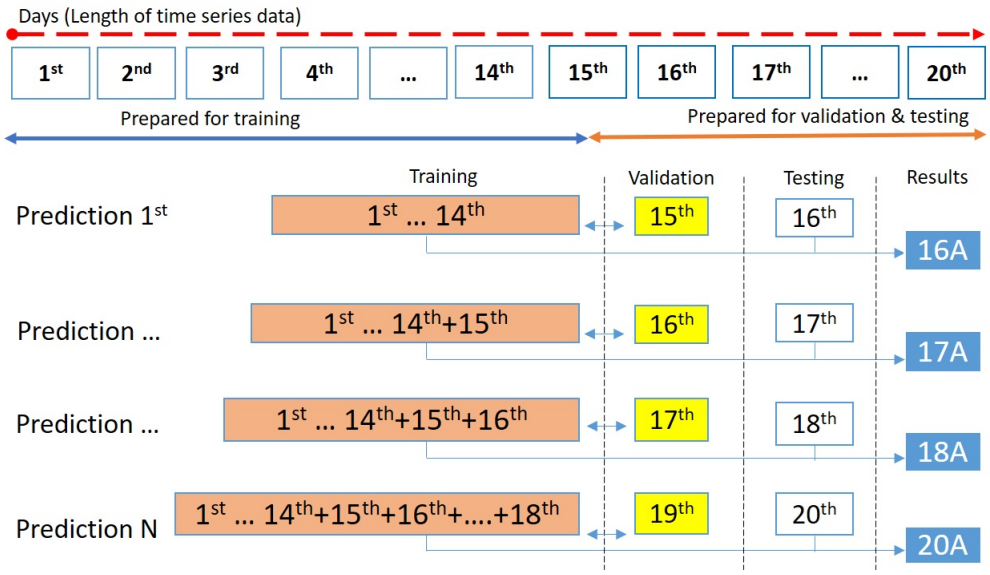

FIGURE 7.3: Training data and prediction iteration

instead of just adding the validation data, I implemented a moving window model such that the length of training data remains the same 14 days in each iteration. This model changes the training data by moving one day forward at each iteration so that each data set has the same number of days of training data, namely 14. Accordingly, the validation data in the previous step was added into the training data in the current step, but the earliest day's data in the training data in the previous step was removed.

In summary, the difference between the proposed models in this study and the model proposed in our previous study [5] is that the previous study model just adds both the predicted result and the validation data in the previous step into the training data in the current step, but the proposed model in this study replaces the earliest day's data in the training data with the validation data in the current step for the next step using the moving window model. Figure 6.1 illustrates the experimental procedure of the proposed model in the training process.

7.3.2 Measuring the Performance of Both Models

DM note: the two figures in Fig 7.4 have the same label "Prediction Error for Inbound Direction" at the top. The bottom one of these should read "...Outbound...".

First, I observed the prediction results for five days to ascertain whether the training data has any significant effect. For the inbound direction on Monday in morning peak (MP), late morning (LM), early afternoon (EA), afternoon peak (AP) and evening (E) the proposed model obtained the lowest prediction error of approximately 8.94% and on Wednesday in morning peak (MP), midday (MD), early afternoon (EA), afternoon peak (AP) and evening (E)the proposed model obtained the lowest prediction error of approximately 9.03%. Next on Tuesday, Thursday and Friday always in the early afternoon (EA) the ANN model obtained the lowest prediction errors of approximately 6.94%, 6.57% and 6.31%, respectively. For the outbound direction, The RF model obtained the lowest prediction error for midday (MD) on Monday of approximately 5.46%, and the SVR model was next, obtaining the lowest prediction error for several days, namely Tuesday in morning peak (MP), Wednesday in afternoon peak (AP), Thursday in morning peak (MP) and Friday in late morning (LM), with values of approximately 5.55%, 7.37%, 5.02% and 6.69%, respectively.

The observations indicate that there is no significant difference in the distribution of MAPE among the three models. However, it can be clearly seen for the inbound direction that the ANN model outperformed the SVR and RF models for the prediction of travel time over several days, especially in time periods with recurrent variability i.e., early afternoon (EA). On the other hand, for the inbound and outbound directions the SVR and RF models give better prediction results compared to the ANN model in predicting travel time in time periods with non-recurrent variability i.e., in morning peak

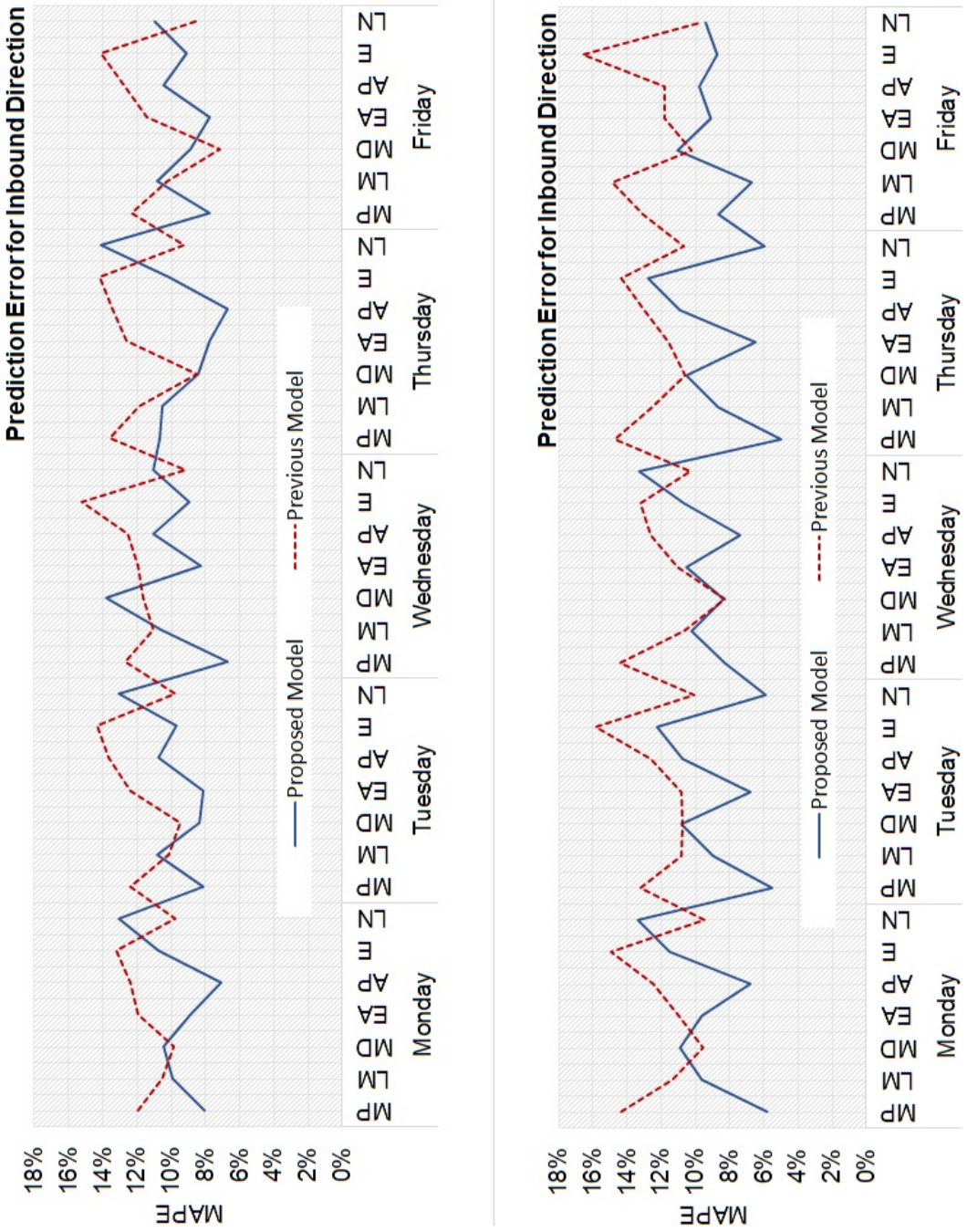

FIGURE 7.4: Comparison of prediction performance between proposed and previous models

(MP), afternoon peak (AP) and evening (E). Recurrent and non-recurrent variability of travel time were observed as described in Section 4.5.2.

For the second measurement, I compared the results achieved by our proposed model and previous model using average MAPE values. The experimental results shown in Figure 7.4 illustrate that our proposed model obtained lower MAPE results than our previous model, except for the two time periods i.e., midday (MD) and late night (LN) for the inbound direction and midday (MD) for the outbound direction. In addition, when focusing on the time periods with non-recurrent properties, morning peak (MP), late morning (LM), afternoon peak (AP), and evening (E), for the inbound and the outbound directions our proposed model outperformed the previous model.

Next, I conducted a t-test from the SPSS software package to see if there are any significant differences between the prediction results. In this procedure, using a paired samples t-test, I compared the means of MAPE values between the proposed model and previous model in each time period for all days.

TABLE 7.2: Paired sample test between the proposed and previous models

Proposed vs	Inbound		Outbound	
Previous	**T-value**	**P-value**	**T-value**	**P-value**
Pair 1 MP	-8.427	**.001**	-7.829	**.001**
Pair 2 LM	-.647	.553	-2.367	.077
Pair 3 MD	1.113	.328	1.633	.178
Pair 4 EA	-13.083	**.000**	-3.270	**.031**
Pair 5 AP	-3.847	**.018**	-3.924	**.017**
Pair 6 E	-7.039	**.002**	-3.567	**.023**
Pair 7 LN	6.057	**.004**	-.270	.801

Table 7.2 shows that there are significant differences between the proposed and previous models. In the inbound direction significant differences are found in five time periods, namely morning peak (MP), early afternoon

(EA), afternoon peak (AP), evening (E) and late night (LN). For the outbound direction, there are four such time periods i.e., morning peak (MP), early afternoon (EA), afternoon peak (AP) and evening (E). The above results show that the proposed model in this study outperformed our previous model and that there are significant differences between them.

7.4 Summary

It is a challenge to predict the bus travel time over each unstable interval between adjacent bus stops accurately, since the variability of the bus travel time has its source in regular variations (recurrent), such as day-to-day variation, and irregular variations (non-recurrent), such as incidents and random variations. Therefore, I conducted two comparison experiments.

First, I determined the impact of using two input variables in our model, namely HATT and DATT. I conducted an experimental comparison between our model and the model proposed by Swardo et al. [74] that only uses HATT as an input variable. Experimental results show that our model provided promising performance in predicting travel time over unstable intervals compared to the method which only uses the HATT data. The results also indicated that our model accurately and dynamically predicted travel time over unstable intervals in each time period in a day. This means that bus travel time can be reasonably estimated when using both DATT and HATT data for unstable intervals with recurrent or non-recurrent variability.

Second, in order to grasp the characteristics of the training data in our previous study [5], I conducted other experiments to predict travel time in each time period for the inbound and the outbound directions according to the training model in our previous study. Since the SVR model had the best performance in this study, I repeated the experiments five times as described

in Section 5.3.2. Experimental results showed that our proposed model provided promising performance in predicting travel time over unstable intervals compared to our previous model [5].

In general, the results indicated that our model accurately and dynamically predicted travel time over unstable intervals in each time period in a day especially in time periods with irregular (non-recurrent) variability in travel time. This means that bus travel time can be reasonably estimated using both DATT and HATT data together over unstable intervals.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

8.1.1 Results and Findings

In this study, bus travel time prediction models were developed using historical bus travel time information. The existing bus travel/arrival time prediction models were studied to decide their limitations. As a result, this study used a heterogeneous (variability) approach to traffic conditions to predict bus travel time over each unstable interval between adjacent bus stops. In the study, I conducted several analyzes of heterogeneous traffic conditions using probe data to understand the real conditions of travel time. In this way, the ability of the models to capture the temporal variations in bus travel time was also determined.

I carried out an exploratory analysis of the travel time variability over each interval between adjacent bus stops on all of the routes to identify key characteristics of travel time. The aim was to generate data having the properties of the observed historical travel time data. This study began the transformation of the historical travel time into time series data by separating components and individually analyzing them to understand the causal mechanisms behind different components that determine the travel time over each interval between adjacent bus stops. Then, I classified the average travel time

in each of eight time periods in a day over days and calculated the average standard deviation of the average travel time over intervals and the standard deviation of the standard deviation of the average travel time over intervals, finally classifying bus travel time over intervals into two categories: stable and unstable.

In the second part of the data analysis, I verified the daily variance of travel time over unstable intervals using a statistical model. In this analysis, I first confirmed that the characteristics of travel time between adjacent bus stops may vary between time periods in a day. Second, I confirmed that the daily variance in travel time tends to be recurrent or non-recurrent, and third, that there are strong correlations of travel time between time periods in a day. These results are in fact the significant factors that influence bus travel time over each unstable interval between adjacent bus stops. Therefore, I employed two types of input data: dynamic average travel time (DATT) in the time period right before the current one and historical average travel time (HATT) in the same time periods over the past several days to build our prediction model of travel time over unstable intervals.

Next, the regular and irregular variability (recurring and non-recurrent patterns) of bus travel time were modeled basically using time series methods based on Artificial Neural Network (ANN), Support Vector Machine Regression (SVR) and Random Forest (RF). In general, the results of the three models showed acceptable performance and a reasonable error range in predicting the travel time over each unstable interval. However, in a comparison of the models, it can be clearly seen for the inbound direction that the ANN model outperformed the SVR and RF models for the prediction of travel time on several days, especially in the time periods with recurrent variability like early afternoon (EA). On the other hand, for the inbound and outbound directions, the SVR and RF models give better prediction results than the ANN

model in predicting travel time in time periods with non-recurrent variability, namely morning peak (MP), afternoon peak (AP) and evening (E).

Since existing stochastic models have not explicitly considered the characteristics of travel time variability between off-peak and peak-hour periods, this study also addresses the problem of predicting bus travel time over unstable intervals influenced by heterogeneous factors between time periods in a day and day to day. Models were built using two types of machine learning techniques, the ANN and SVR methods. In these models, to predict travel time over each unstable interval, I used two schemes, namely to predict travel time in the off-peak periods, using only one input variable: HATT, while using two input variables, DATT and HATT, for peak hour periods. The results were achieved by predicting travel time over unstable intervals both in the peak hours and in the off-peak periods of weekdays. It is observed that, by using two input variables for peak-hour periods and one input variable for off-peak periods, the prediction accuracy can be effectively improved. The results show that the ANN and SVR models capture the periodic variations during peak-hour periods.

8.1.2 Research Contribution

The objective of this research was to develop a travel time prediction algorithm with a focus on unstable intervals between adjacent bus stops. The case study conducted in this research includes travel time variability. Thus, this research conducted a three-stage exploratory analysis of travel time variability before building a prediction model.

First, the daily average of travel time for a month was observed, and it became clear that the average travel time may vary by up to 100% between time periods in a day and over days. Using the average standard deviation of the average travel time over intervals and the standard deviation of the standard

deviation of the average travel time over intervals, I succeeded in classifying all intervals into stable and unstable ones. Then, I further subdivided each of the two types into three sub-classes: weak, medium and strong. This is an important contribution to research which aims to predict bus travel time under heterogeneous conditions in the absence of data concerning variables such as traffic and weather.

Second, focusing on each unstable interval, I confirmed the travel time variability over days. In this stage, using statistical analysis, I compared the average travel time over intervals for the eight time periods in a day and for the same time periods over days. The results show that there are significant differences between the average travel time over days and that the characteristics of travel time over each unstable interval between adjacent bus stops may vary between time periods in a day and over days. The analysis of comparisons between the same time periods over days shows the following: in weekday peak-hour periods, namely morning peak (MP), late morning (LM), afternoon peak (AP) and evening (E), the bus travel time may significantly increase due to unexpected events such as heavy traffic volume, accidents, road construction or weather; in the off-peak periods, namely early morning (EM), midday (MD), early afternoon (EA) and late night (LN), the travel times in the same time period are fairly constant for weekdays. These results show that the variance in travel time in peak-hour periods between days is consistently higher. This is an important finding because it indicates that it is insufficient to only use the variance of travel time among the eight time periods in a day to predict the travel time.

Finally, about the correlation of travel time between adjacent time periods in a day over days, the results show that there are strong or moderate correlations of the average travel time over each unstable interval between time periods in a day, in particular when two time periods are near to each other. This is also an important finding indicating that travel times in the previous

time periods are a useful factor in predicting travel times in the later time periods and in building a predictive model for travel times.

Furthermore, from the results, I proposed a method which built nonlinear dynamical models to predict bus travel time over each unstable interval between adjacent bus stops in a day. The inclusion of the travel time variability into the prediction model is an important contribution to the research efforts to predict bus travel times.

A significant attempt has been made in this research to explain the phenomenon of travel time variability included in the travel time prediction model in the absence of traffic condition data and weather data. It was shown that the proposed model was able to produce significant improvements in accurately predicting bus travel time over each unstable interval under heterogeneous conditions. This means that bus travel time can be reasonably estimated over each unstable interval using both DATT and HATT data together. Next, for the second model, which uses the input variables selectively with a distinction between off-peak and peak-hour periods, the prediction performance of the model showed an acceptable prediction performance.

8.2 Recommendations for Future Research

As with any data mining techniques, this study could benefit from a larger sample size of historical travel times. The observational sample size requirements were fulfilled, but a larger sample size could be used for each of the analyses performed. This could include more data about the speed of buses, dwell time, traffic conditions and weather on the road. The models used in this thesis should be easily implementable for other conditions where the prediction of bus travel time is implemented.

In addition, a larger sample size of historical travel times can obviously increase the accuracy of the analysis of the travel time characteristics. The

literature review showed that a larger sample size creates an opportunity to use more complex processes to obtain better estimates of the probabilities.

Further building on this study, future research should also be conducted during different peak-hour periods in a day over intervals between two adjacent bus stops, or be expanded to consider different characteristics between routes in urban areas and those in rural areas. Information concerning the dwell time of buses at bus stops should also be taken into account by further analysis, including through study using non-linear regression before making a travel time prediction model. Such results will provide a more robust understanding of what occurs between bus stops, between time periods in a day and the variance between buses in dwell time at bus stops.

References

[1] Abma, R.: Assessing Travel Time Reliability in Urban Networks from a Road User Perspective. Ph.D. thesis, Citeseer (2014)

[2] Amita, J., Jain, S., Garg, P.: Prediction of bus travel time using ann: A case study in delhi. Transportation Research Procedia 17, 263–272 (2016)

[3] As, M., Mine, T.: Empirical study of travel time variability using bus probe data. In: Agents (ICA), IEEE International Conference on. pp. 146–149. IEEE (2016), http://doi.ieeecomputersociety.org/10.1109/ICA.2016.050

[4] As, M., Mine, T.: An adaptive approach for predicting bus travel time over unstable intervals. In: The 16th ITS Asia-Pacific Forum FUKUOKA 2018. p. 146. ITS AP Forum Fukuoka (2018)

[5] As, M., Mine, T.: Dynamic bus travel time prediction using an ann-based model. In: The 12th International Conference on Ubiquitous Information Management and Communication,(IMCOM). p. 8pages. ACM (2018), http://dx.doi.org/10.1145/3164541.3164630

[6] As, M., Mine, T., Yamaguchi, T.: Prediction of bus travel time over unstable intervals between two adjacent bus stops. International Journal of Intelligent Transportation Systems Research pp. 1–12 (2018), http://dx.doi.org/10.1007/s13177-018-0169-3

[7] As Mansur, Hiroyuki, N., Mine, T.: Estimation of travel time variability using bus probe data. In: 6th IEEE International Conference

on Advanced Logistics and Transport (ICALT). pp. 68–74. IEEE (2017), http://dx.doi.org/10.1109/ICAdLT.2017.8547006

[8] Asensio, J., Matas, A.: Commuters' valuation of travel time variability. Transportation Research Part E: Logistics and Transportation Review 44(6), 1074–1085 (2008)

[9] Bai, C., Peng, Z.R., Lu, Q.C., Sun, J.: Dynamic bus travel time prediction models on road with multiple bus routes. Computational intelligence and neuroscience 2015, 63 (2015)

[10] Basak, D., Pal, S., Patranabis, D.C.: Support vector regression. Neural Information Processing-Letters and Reviews 11(10), 203–224 (2007)

[11] Bin, Y., Zhongzhen, Y., Baozhen, Y.: Bus arrival time prediction using support vector machines. Journal of Intelligent Transportation Systems 10(4), 151–158 (2006)

[12] Börjesson, M., Eliasson, J., Franklin, J.P.: Valuations of travel time variability in scheduling versus mean–variance models. Transportation Research Part B: Methodological 46(7), 855–873 (2012)

[13] Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time series analysis: forecasting and control. John Wiley & Sons (2015)

[14] Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (Oct 2001), https://doi.org/10.1023/A:1010933404324

[15] Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2(3), 27 (2011)

[16] Chang, H., Park, D., Lee, S., Lee, H., Baek, S.: Dynamic multi-interval bus travel time prediction using bus transit data. Transportmetrica 6(1), 19–38 (2010)

[17] Chen, C., Skabardonis, A., Varaiya, P.: Travel-time reliability as a measure of service. Transportation Research Record: Journal of the Transportation Research Board 1(1855), 74–79 (2003)

[18] Chen, G., Yang, X., Zhang, D., Teng, J.: Historical travel time based bus-arrival-time prediction model. In: ICCTP 2011: Towards Sustainable Transportation Systems, pp. 1493–1504. ASCE Library (2011)

[19] Chen, M., Liu, X., Xia, J., Chien, S.I.: A dynamic bus-arrival time prediction model based on apc data. Computer-Aided Civil and Infrastructure Engineering 19(5), 364–376 (2004)

[20] Chen, S., Billings, S., Grant, P.: Non-linear system identification using neural networks. International journal of control 51(6), 1191–1214 (1990)

[21] Chien, S.I.J., Ding, Y., Wei, C.: Dynamic bus arrival time prediction with artificial neural networks. Journal of Transportation Engineering 128(5), 429–438 (2002)

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

[23] CONTAIN, T.A.: Real-time travel time estimates using media access control address matching. ITE JOURNAL (2008)

[24] Cootes, T.F., Ionita, M.C., Lindner, C., Sauer, P.: Robust and accurate shape model fitting using random forest regression voting. In: European Conference on Computer Vision. pp. 278–291. Springer (2012)

[25] Demuth, H., Beale, M., Hagan, M.: Neural network toolbox^{TM} 6. User's guide 10, 11 (2008)

[26] Diaconescu, E.: The use of narx neural networks to predict chaotic time series. Wseas Transactions on computer research 3(3), 182–191 (2008)

[27] Eliasson, J.: Forecasting travel time variability. In: European Transport Conference (2006)

[28] Engelson, L., Fosgerau, M.: Additive measures of travel time variability. Transportation research part B: methodological 45(10), 1560–1571 (2011)

[29] Fan, W., Gurmu, Z.: Dynamic travel time prediction models for buses using only gps data. International Journal of Transportation Science and Technology 4(4), 353–366 (2015)

[30] Feng, W., Figliozzi, M., Bertini, R.L.: Quantifying the joint impacts of stop locations, signalized intersections, and traffic conditions on bus travel time. Public Transport 7(3), 391–408 (2015)

[31] Fosgerau, M., Engelson, L.: The value of travel time variance. Transportation Research Part B: Methodological 45(1), 1–8 (2011)

[32] Fosgerau, M., Fukuda, D.: Valuing travel time variability: Characteristics of the travel time distribution on an urban road. Transportation research part c: emerging technologies 24, 83–101 (2012)

[33] Fosgerau, M., Hjorth, K., Brems, C., Fukuda, D.: Travel time variability: Definition and valuation. DTU Transport (2008)

[34] Gal, A., Mandelbaum, A., Schnitzler, F., Senderovich, A., Weidlich, M.: Traveling time prediction in scheduled transportation with journey segments. Information Systems 64, 266–280 (2017)

[35] Gurmu, Z.K., Fan, W.D.: Artificial neural network travel time prediction model for buses using only gps data. Journal of Public Transportation 17(2), 3 (2014)

[36] Hamner, B.: Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow. In: Data Mining

Workshops (ICDMW), 2010 IEEE International Conference on. pp. 1357–1359. IEEE (2010)

[37] Hensher, D.A.: Identifying the influence of stated choice design dimensionality on willingness to pay for travel time savings. Journal of Transport Economics and Policy (JTEP) 38(3), 425–446 (2004)

[38] Janacek, G.: Time series analysis forecasting and control. Journal of Time Series Analysis 31(4), 303–303 (2010)

[39] Jenelius, E.: The value of travel time variability with trip chains, flexible scheduling and correlated travel times. Transportation Research Part B: Methodological 46(6), 762–780 (2012)

[40] Jeong, R., Rilett, R.: Bus arrival time prediction using artificial neural network model. In: Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on. pp. 988–993. IEEE (2004)

[41] Kho, S., Cho, J.: Estimating average travel times from bus travel times. Proc. East. Asia Soc. Transp. Stud 3(2), 45–55 (2001)

[42] Khosravi, A., Mazloumi, E., Nahavandi, S., Creighton, D., Van Lint, J.: A genetic algorithm-based method for improving quality of travel time prediction intervals. Transportation Research Part C: Emerging Technologies 19(6), 1364–1376 (2011)

[43] Kumar, S.V., Vanajakshi, L.: Urban arterial travel time estimation using buses as probes. Arabian Journal for Science and Engineering 39(11), 7555–7567 (2014)

[44] Kwon, J., Coifman, B., Bickel, P.: Day-to-day travel-time trends and travel-time prediction from loop-detector data. Transportation Research Record: Journal of the Transportation Research Board 1717, 120–129 (2000)

[45] Lee, W.C., Si, W., Chen, L.J., Chen, M.C.: Http: a new framework for bus travel time prediction based on historical trajectories. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems. pp. 279–288. ACM (2012)

[46] Leshem, G., Ritov, Y.: 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 (2007)

[47] Li, C.S., Chen, M.C.: Identifying important variables for predicting travel time of freeway with non-recurrent congestion with neural networks. Neural Computing and Applications 23(6), 1611–1629 (2013)

[48] Li, R.: Enhancing motorway travel time prediction models through explicit incorporation of travel time variability. Ph.D. thesis, Monash University (2006)

[49] Li, R., Rose, G.: Incorporating uncertainty into short-term travel time predictions. Transportation Research Part C: Emerging Technologies 19(6), 1006–1018 (2011)

[50] Li, R., Rose, G., Sarvi, M.: Using automatic vehicle identification data to gain insight into travel time variability and its causes. Transportation Research Record 1945(1), 24–32 (2006)

[51] Liaw, A., Wiener, M., et al.: Classification and regression by randomforest. R news 2(3), 18–22 (2002)

[52] Liu, H., Van Zuylen, H.J., Van Lint, H., Chen, Y., Zhang, K.: Prediction of urban travel times with intersection delays. In: Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE. pp. 402–407. IEEE (2005)

[53] Ma, Z., Ferreira, L., Mesbah, M., Zhu, S.: Modeling distributions of travel time variability for bus operations. Journal of Advanced Transportation 50(1), 6–24 (2016)

[54] Martchouk, M., Mannering, F.L., et al.: Analysis of travel time reliability on indiana interstates. Tech. rep., NEXTRANS (2009)

[55] Mazloumi, E., Currie, G., Rose, G.: Using traffic flow data to predict bus travel time variability through an enhanced artificial neural network. In: World Congress on Transport Research, 12th, 2010. vol. 3377 (2010)

[56] Mazloumi, E., Rose, G., Currie, G., Moridpour, S.: Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction. Engineering Applications of Artificial Intelligence 24(3), 534–542 (2011)

[57] Menezes Jr, J.M.P., Barreto, G.A.: Long-term time series prediction with the narx network: an empirical evaluation. Neurocomputing 71(16-18), 3335–3343 (2008)

[58] Moreira, J.P.C.L.M.: Travel Time Prediction for the Planning of Mass Transit Companies: a Machine Learning Approach. Ph.D. thesis, Universidade do Porto (2008)

[59] Müller, K.R., Smola, A., Rätsch, G., Schölkopf, B., Kohlmorgen, J., Vapnik, V.: Using support vector machines for time series prediction. Advances in kernel methods—support vector learning pp. 243–254 (1999)

[60] Oh, J.S., Chung, Y.: Calculation of travel time variability from loop detector data. Transportation Research Record: Journal of the Transportation Research Board 12(1945), 12–23 (2006)

[61] Padmanaban, R., Vanajakshi, L., Subramanian, S.C.: Estimation of bus travel time incorporating dwell time for apts applications. In: Intelligent Vehicles Symposium, 2009 IEEE. pp. 955–959. IEEE (2009)

[62] Park, D., Rilett, L.R., Han, G.: Spectral basis neural networks for real-time travel time forecasting. Journal of Transportation Engineering 125(6), 515–523 (1999)

[63] Patnaik, J., Chien, S., Bladikas, A.: Estimation of bus arrival times using apc data. Journal of public transportation 7(1), 1 (2004)

[64] Peer, S., Koopmans, C.C., Verhoef, E.T.: Prediction of travel time variability for cost-benefit analysis. Transportation Research Part A: Policy and Practice 46(1), 79–90 (2012)

[65] Ramakrishna, Y., Ramakrishna, P., Lakshmanan, V., Sivanandan, R.: Bus travel time prediction using gps data. Proceedings Map India (2006)

[66] Ramezani, M., Geroliminis, N.: On the estimation of arterial route travel time distribution with markov chains. Transportation Research Part B: Methodological 46(10), 1576–1590 (2012)

[67] Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M.: Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews 71, 804–818 (2015)

[68] Sapankevych, N.I., Sankar, R.: Time series prediction using support vector machines: a survey. IEEE Computational Intelligence Magazine 4(2) (2009)

[69] Satoshi, F.: On the subject of probe data. Director of the Research Center for Advanced Information Technology (2011)

[70] Shalaby, A., Farhan, A.: Bus travel time prediction model for dynamic operations control and passenger information systems. Transportation Research Board 2 (2003)

[71] Shalaby, A., Farhan, A.: Prediction model of bus arrival and departure times using avl and apc data. Journal of Public Transportation 7(1), 3 (2004)

[72] Shmueli, G., et al.: To explain or to predict? Statistical science 25(3), 289–310 (2010)

[73] Susilawati, S., Taylor, M.A., Somenahalli, S.V.: Distributions of travel time variability on urban roads. Journal of Advanced Transportation 47(8), 720–736 (2013)

[74] Suwardo, W., Napiah, M., Kamaruddin, I.: Arima models for bus travel time prediction. Journal of the Institute of Engineers Malaysia pp. 49–58 (2010)

[75] Syrjärinne, P.: Urban Traffic Analysis with Bus Location Data. Ph.D. thesis, Tampere University Press (2016)

[76] Tyralis, H., Papacharalampous, G.: Variable selection in time series forecasting using random forests. Algorithms 10(4), 114 (2017)

[77] Uno, N., Kurauchi, F., Tamura, H., Iida, Y.: Using bus probe data for analysis of travel time variability. Journal of Intelligent Transportation Systems 13(1), 2–15 (2009)

[78] Van Lint, J., Van Zuylen, H.: 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 1917, 54–62 (2005)

[79] Vanajakshi, L., Rilett, L.R.: Support vector machine technique for the short term prediction of travel time. In: Intelligent Vehicles Symposium, 2007 IEEE. pp. 600–605. IEEE (2007)

[80] Vanajakshi, L., Subramanian, S.C., Sivanandan, R.: Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses. IET intelligent transport systems 3(1), 1–9 (2009)

[81] Vapnik, V.: The nature of statistical learning theory. Springer science & business media (2013)

[82] Viviano, D.: Bus travel time analysis using real-time data. Ph.D. thesis, LUISS Guido Carli (2016)

[83] Wang, L., Zuo, Z., Fu, J.: Bus arrival time prediction using rbf neural networks adjusted by online data. Procedia-Social and Behavioral Sciences 138, 67–75 (2014)

[84] Washington, S.P., Karlaftis, M.G., Mannering, F.: Statistical and econometric methods for transportation data analysis. Chapman and Hall/CRC (2010)

[85] Wu, C.H., Ho, J.M., Lee, D.T.: Travel-time prediction with support vector regression. IEEE transactions on intelligent transportation systems 5(4), 276–281 (2004)

[86] Wu, X., Zhang, H.: Analysis of time-dependent travel time reliability for urban corridors: A cast study in houston. In: Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on. pp. 1939–1944. IEEE (2016)

[87] Yassin, I.M., Taib, M.N., Adnan, R.: Recent advancements & methodologies in system identification: A review. Scientific Research Journal 1(1), 14–33 (2013)

[88] Yetiskul, E., Senbil, M.: Public bus transit travel-time variability in ankara (turkey). Transport Policy 23, 50–59 (2012)

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

[90] YU, B., LU, J., YU, B., YANG, Z.: An adaptive bus arrival time prediction model. Journal of the Eastern Asia Society for Transportation Studies 8, 1126–1136 (2010)

[91] Zhang, F., Zhu, X., Hu, T., Guo, W., Chen, C., Liu, L.: Urban link travel time prediction based on a gradient boosting method considering spatiotemporal correlations. ISPRS International Journal of Geo-Information 5(11), 201 (2016)

[92] Zhang, X., Rice, J.A.: Short-term travel time prediction. Transportation Research Part C: Emerging Technologies 11(3-4), 187–210 (2003)

Appendix A

Publish Work

A.1 Journal Paper and Book:

Prediction of Bus Travel Time Over Unstable Intervals between TwoAdjacent Bus Stops. International Journal of Intelligent Transportation Systems Research. http://dx.doi.org/10.1007/s13177-018-0169-3

Springer book "Intelligent Transport Systems for Everyone's Mobility" to be published by Springer.

A.2 Conference Papers:

Mansur AS, Tsunenori Mine. Empirical Study of Travel Time Variability Using Bus Probe Data. In: Agents (ICA), IEEE International Conference on. IEEE, 2016. p. 146-149.
http://doi.ieeecomputersociety.org/10.1109/ICA.2016.050

Mansur AS, Tsunenori Mine. Estimation of travel time variability using bus probe data. In: 6th IEEE International Conference on Advanced Logistics and Transport (ICALT). IEEE 2017, pp.68–74.
http://dx.doi.org/10.1109/ICAdLT.2017.8547006

Mansur AS, Tsunenori Mine. Dynamic Bus Travel Time Prediction Using an ANN-based Model. In: Proceedings of the 12th International

Conference on Ubiquitous Information Management and Communication. ACM, 2018. p. 8pages.
http://dx.doi.org/10.1145/3164541.3164630

Mansur AS, Tsunenori Mine. An adaptive approach for predicting bus travel time over unstable intervals. In: The 16th ITS Asia-Pacific Forum FUKUOKA. ITS Asia Pacific and ITS Japan 2018, pp. 146–160
https://www.itsap-fukuoka.jp