Bus Arrival Time Reliability Analyses and Dynamic Prediction Model Based on Multi-source Data

by

Ling Shi

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

In

Transportation Engineering

Department of Civil and Environmental Engineering

University of Alberta

Abstract

In recent years, with increased urbanization and vehicle ownership, congestion levels have been increasing in urban areas although new infrastructure is being developed to meet the increasing demand. An understanding of bus service reliability is necessary to develop strategies that help transit agencies provide better services. In addition, a reliable and accurate vehicle arrival prediction system can help in making public transportation more attractive. This study first aims to quantify and determine temporal and weather related bus arrival time variability in Edmonton. A multinomial logit model is developed and estimated, which relates early, late and on-time bus arrivals to weather, temporal and operating characteristics. The model results show that the probability of on-time failures increases during PM peak periods, as buses progress further along their routes and under adverse weather conditions. Secondly, a proposed bus tracing algorithm and bus arrival time prediction algorithm are then applied to predict bus travel time using Global Positioning System (GPS) data and Vehicle Detect System (VDS) data, also, a regression model based on the factors found from the multinomial logit model is applied as a comparison. A case study is conducted on one selected bus routes in Edmonton, to evaluate the performance of the proposed algorithm in terms of prediction accuracy. The results indicate that the proposed algorithm is capable of achieving satisfactory accuracy in predicting bus arrival time.

Acknowledgement

This thesis could not have been completed without the help and support of many people, only a few of whom are listed below.

I would like to thank my committee members for their guidance. I am especially grateful to my supervisor, Dr. Qiu, who has always encouraged me to do my best during my MSC program at the University of Alberta.

Thanks to my team members in the Centre for Smart Transportation which has a wonderful research and collaboration atmosphere. I would like to thank Dr. Liu, Dr. Zhang and Jiangcheng Li. Their instruction provides significant support not only on the research ideas, but also on the research techniques. I also would like to express thanks to the City of Edmonton for providing data for this study. The contents of this paper reflect the views of the authors and not necessarily the view of the City of Edmonton.

Finally, I wish to express appreciation to my parents and girlfriend. Only with their support and understanding can I finally finish my graduation study. Thanks for their loving consideration and great confidence in me through all these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List OF ABBREVIATION

AFC	Automatic Fare Collection
ANN	Artificial Neural Networks
APC	Automatic Passenger Counters
APTS	Advanced Public Transportation Systems
AVL	Automatic Vehicle Location
AVT	Automatic Vehicle Track
DATS	Disabled Adult Transit Service
ETS	Edmonton Transit System
GIS	Geographic Information System
GPS	Global Position System
GTFS	General Transit Feed Specification
ITS	Intelligent Transportation Systems
LRS	Linear Referencing System
LRT	Light Rail Train
MAPE	Mean Absolute Error
PA	Percentage of Arrivals within On-Time Threshold
RMSE	Root Mean Squared Error
SD	Standard Deviation of Arrival Time Variance
TSR	Transit Service Reliability
VDS	Vehicle Detect System

CHAPTER 1 INTRODUCTION

1.1 Background

In recent years, congestion has become a big problem and deteriorated the quality of life of people in many developed and developing countries. This is mainly due to many factors, especially the increasing number of vehicle ownerships. In the developed countries such as United States, a report from the ICF Consulting found that, in 2003, about 76% of Americans chose privately owned vehicles for their commute to work (Consulting 2003), and around 79.5% of them drive alone when commuting (McKenzie and Rapino 2009). Same trend in the developing countries, a report from the Beijing Transportation Research Center found that 34% of those living in Beijing chose their privately owned vehicles for commuting in 2010 (Beijing Transportation Research Center 2010). As a result, the high percentage of using privately owned vehicles will increase congestion. While congestion will lead to many problems such as increase in energy consumption, air pollution and travel time. In order to relieve congestion, different approaches have been adopted in both demand side and supply side. The approaches in demand side mainly focus on a more efficiently way to use the existing system. The approaches in supply sides mainly concentrate on infrastructure expansion. However, the approaches in supply sides such as infrastructure expansion cannot meet the vehicle population growth rate, and hence, more solutions in demand side need to be explored. In this regard, Intelligent Transportation Systems (ITS) is considered as a good demand side approach which involves many functional areas, including Advanced Public Transportation Systems (APTS).

Public transit has been recognized as the key to developing the future environmentally conscious and sustainable transportation system. A report from the Texas Transportation Institute

indicated that the public transit saved 27% for travel delay in 2003, which is equal to more than 1.1 billion hours of travel time (Texas Transportation Institute 2005). Therefore, it is of critical importance to improve the efficiency of Transit Service Reliability (TSR). While TSR can be affected by many factors, such as traffic conditions, signal delay at traffic lights, road geometry, uncertain passenger demand, driver behavior, vehicle accidents, bus stop locations and weather. By investigating and quantifying the impacts of TSR, transit operators can prioritize measures or investments that tackle the main sources of unreliable transit service, and propose appropriate strategies to improve TSR (Crout 2007). For individuals, bus arrival time is more concern for them. The provision of accurate bus arrival time prediction is important to attract additional ridership and increase the satisfaction of transit users (Hensher, Stopher, and Bullock 2003; Murray and Wu 2003).

In recent years, many transit agencies have used a range of advanced technologies to collect amounts of multi-source data, such as Automatic Passenger Counters (APC) data, Automatic Vehicle Location (AVL) data, General Transit Feed Specification (GTFS) data, weather data and smart bus GPS data. With advances in information technologies in ITS, the availability of public transit data has been increasing in the past decades, which has gradually shifted the public transit system into a data-rich environment, and makes it realizable to do further TSR analysis and predict bus arrival time more preciously.

1.2 Problem Statement

TSR may be affected by a number of internal and external factors, such as passenger demand and its distribution along the route, traffic conditions (recurrent traffic congestion and non-recurrent incidences), driving habits, weather conditions, transit operations as well as

network design. If the impacts of TSR can be investigated and quantified, transit operators can prioritize measures or investments that tackle the main sources of unreliable transit service, and propose appropriate strategies to improve TSR (Crout 2007). In addition, providing accurate information about bus arrival time to passengers is an attractive popular application. It will be useful to passengers to reduce the waiting time at bus stops or to make reasonable travel arrangements before making a trip. However, for this to be effective, the information provided to passengers should be reliable and accurate.

With advances in new technology systems and data integration initiatives, public transit agencies are gaining increasing amounts of multi-source data, such as APC, AVL, smart cards, weather information and collision data. These rich data sources have the potential to generate much more and better information about the transit system and customers for decision making, such as much more and better information to analysis the factors affecting the bus on-time reliability and improve the bus arrival time prediction. However, one sufficient condition is to understand how to derive value from these data. Nowadays, one typical case is that big data are available but have not yet been harnessed by users and operators. In addition, those multi-source data are maintained by different branches with diversity of formats, and the transit data is location sensitive, so in order to use those multi-source data, the data integration is needed as well as the transit data need to be projected onto the underlying transit network. Before, lots of bus on-time reliability analyzes are only based on one dimension data source, while with the achieving of multi-source data, this problem can be extended by using those integrated multi-source data. Furthermore, nowadays, lots of transportation agencies have installed GPS sensors in buses, so the real time bus GPS data can be integrated with those multi-source data to improve the bus arrival time prediction.

1.3 Research Scope

Using the Edmonton Transit System (ETS) as a case study, this study aims to investigate the use of multiple data sources to quantify bus on-time performance and predict bus arrival time. There are three specific goals of this thesis:

a) Extract the transit related information from APC data, GTFS data, and smart bus GPS data as well as integrate the extracted transit data with the road network.
b) Analyze the impacts of weather, temporal and operating characteristics affecting bus on-time reliability and predict the bus arrival time based on those factors.
c) Develop a bus arrival time prediction model to predict the bus arrival time based on the smart bus real-time GPS data and loop detector data.

1.4 Structure of Thesis

The structure for this thesis is as follows:

Chapter 1 introduces the background of TSR and bus arrival time prediction as well as the problem statement and the research scopes.

Chapter 2 is the literature review chapter, which discusses about the works that have been done in past few years for the evaluation of factors affecting bus on-time performance and the bus arrival time prediction.

Chapter 3 introduces the data preprocessing of multiple data sources for the evaluation of factors affecting bus on-time performance and the bus arrival time prediction.

Chapter 4 describes the evaluation of factors affecting bus on-time performance using multi-source data.

Chapter 5 describes the bus arrival time prediction based on the real-time GPS and loop detector data and the bus arrival time prediction based on the factors affecting bus on-time performance found in chapter 4.

Chapter 6 presents the conclusion and provides suggestions for the future works.

CHAPTER 2 LITERATURE REVIEW

2.1 Transit Service Reliability

TSR is a key performance index to evaluate service quality, as it is a vital issue affecting transit's status as a desired alternative to private transportation (Arhin and PTOE 2013; J., Liu, and Yang 2014). In 2003, ICF Consulting present a report that around 76% of Americans chose privately owned vehicles for their commute to work (Consulting 2003), and 79.5% of them drive alone when commuting(McKenzie and Rapino 2009). Also, Beijing Transportation Research Center found that 34% of those living in Beijing chose their privately owned vehicles for commuting in 2010 (Beijing Transportation Research Center 2010). As a result, the high possibility of using privately owned vehicles will increase air pollution, energy consumption and congestion. While, public transit is an effective measure to avoid such results, for example, a report from the Texas Transportation Institute indicated that the public transit saved 27% for travel delay in 2003, which is equal to more than 1.1 billion hours of travel time (Texas Transportation Institute 2005). Therefore, it is of critical importance to improve the efficiency of TSR.

Typically, TSR is usually measured in terms of schedule adherence, running times, headways, and passenger waiting times (Boilé 2001; Herbert S. Levinson 2005). Schedule adherence is calculated as actual departure time minus scheduled departure time, and is consistently ranked as one of the major concerns of passengers, especially for low frequency routes. Passengers will arrive at their destinations late due to the poor schedule adherence, and this can finally cause customer satisfaction issues (Cham 2006). While on-time performance is a common way to measure schedule adherence, and it is defined as the percentages of buses that

depart from a location with a time window in advance (Kittelson & Associates 2013). The standard of 'on time' is defined as a bus departures or arrives no more than one minute early and five minutes late by most of transit agencies (Bates 1986). Another measure of TSR is headway, which is calculated as actual headway minus scheduled headway, and is more focus on spacing between buses. If the headway value is negative, it means a bus is falling behind its leader and vice versa. The negative of headway value will cause additional delay, and result as more waiting time and more passenger loads (Cham 2006). Running time is also an important measure of TSR, which is calculated as actual running time minus scheduled running time. Running time is more focus on link level, and if the value of running time is positive, it means that a bus takes more time to traverse that link (Herbert S. Levinson 2005).

2.2 Transit Schedule Adherence Analysis

The on-time performance can be affected by a number of factors, such as passenger demand and its distribution along the route, traffic conditions (recurrent traffic congestion and non-recurrent incidences), driving habits, weather conditions, transit operations as well as network design (Sterman and Schofer 1976; Kjmpel 2001). The literature review is conducted in terms of weather factors, temporal factors and operational factors.

2.2.1 Weather Factors

Weather condition is one of the most important factors that are highly associated with bus on-time performance. Mesbah et al. investigated the effect of seasonal variation of daylight hours on the reliability of public transport service. A linear regression model was developed to regress tram travel time on schedule travel time, time difference between 5:00 am to sunrise and time difference between 5:00 to trip start time. Results indicated that daylight start time has a small

but statistically significant effect on service travel time (Mesbah, Currie, and Prohens 2014). In addition, they investigated the effect of weather conditions on the travel time reliability of fifteen randomly selected Melbourne streetcar (tram) routes. Ordinary least square regression analysis was conducted to regress travel time on various weather effects. The results indicated that only precipitation and air temperature are significant in their effect on tram travel time (Mesbah, Lin, and Currie 2015). Levinson and Cham found that the inclement weather can cause slower driving speeds, and the bus travel time will be extended (H S Levinson 1991; Cham 2006). Also, Tétreault et al., El-Geneidy et al. and Diab et al. found that weather has a significant impact on both bus dwell time and bus travel time (Tétreault and El-Geneidy 2010; E. I. Diab and El-Geneidy 2012; A.M. El-Geneidy, Horning, and Krizek 2011).

2.2.2 Operational Factors

Travel distance is one of the most important internal factors that are highly associated with bus on-time performance. Chen et al. investigated the bus service reliability in Beijing, China, and found a high correlation between service reliability and route length, headway, distance from the stop to the origin terminal, and the provision of exclusive bus lanes (Chen et al. 2009). Furthermore, the study from Van Oort et al. has a clear indication that shorter routes tend to be more reliable (Van Oort and Nes 2009).

A few studies have investigated the impact of signalized intersections on bus travel time. The work from El-Geneidy et al. found that bus travel time is extended with an average of 26 seconds by each intersection and each stop sign adds an average of 16 seconds to bus travel time (Ahmed M El-Geneidy, Hourdos, and Horning 2009). For the other study areas, Abkowitz et al., McKnight et al., and Albright et al. found that bus travel time is extended with an average of 8, 11 and 10 seconds by each intersection, respectively (Abkowitz and Engelstein 1984; McKnight

et al. 2004; Albright and Figliozzi 2012). Furthermore, the work of Figliozzi et al. shows that passing through, turning left and turning right at an intersection contributes additional 5, 20 and 38 seconds to bus travel time (Figliozzi and Feng 2012).

The number and spacing of bus stops also has an impact on bus on-time performance. Strathman et al., Tétreault et al., Diab et al. and Slavin et al. found that each additional bus stop contributes 5 to 26 seconds to bus travel time by using the actual number of stops in a trip as a variable (Strathman et al. 2002; Tétreault and El-Geneidy 2010; E. I. Diab and El-Geneidy 2012; Slavin et al. 2013). The studies from McKnight et al., EI-Geneidy et al. and El-Geneidy et al. shows that each additional bus stop gives rise to an average of 5 to 13 seconds increase for bus travel time by using the number of scheduled bus stops as the variable (Ahmed M El-Geneidy, Hourdos, and Horning 2009; McKnight et al. 2004; A.M. El-Geneidy, Horning, and Krizek 2011). Also, El-Geneidy et al. found that the percentage of scheduled bus stops has a positive and significant impact on bus travel time (A.M. El-Geneidy and Surprenant-Legault 2010).

Bus departure delay is analyzed as an internal factor for bus on-time performance by many studies. Strathman et al., El-Geneidy et al. and Figliozziet al. found that bus departure delay has a negative impact on bus travel time (Figliozzi and Feng 2012; Strathman et al. 2002; A.M. El-Geneidy and Surprenant-Legault 2010).

Other internal factors were also found to have significant impacts on bus travel time. For example, bus stop location type also has a significant impact on bus travel time. Albright et al. found that, compared to far-side stops, near-side bus stops can decrease an average of 3.7 seconds in travel time(Albright and Figliozzi 2012). Some studies found that bus route type and bus vehicle type also have a significant impact on bus travel time (Figliozzi and Feng 2012; Strathman et al. 2002; Dueker et al. 2004).

Other external factors were also found to have significant impacts on bus travel time. For example, passenger boarding and alighting activities also have an impact on bus dwell times at bus stops and bus travel time. The work of Abkowitz et al. was among the earliest studies on running time variation, and they found that each passenger boarding and alighting activity contributes 6 and 4 seconds to bus travel time, respectively (Abkowitz and Engelstein 1984). However, the number of bus stops was not included in this study. By adding the number of bus stops, Bertini et al., Dueker et al., Figliozzi et al. and Slavin et al. found that each passenger boarding activity contributes 3.4 to 3.7 seconds to bus travel time, and each passenger alighting activity contributes 0.4 to 1.5 seconds to bus travel time (Figliozzi and Feng 2012; Slavin et al. 2013; Bertini and El-Geneidy 2004; Dueker et al. 2004).

2.2.3 Temporal Factors

Traffic congestion is one of the most important external factors that are highly associated with bus on-time performance. Studies from Levinson et al., Guertset al., Cham and Xuan et al. indicated that traffic congestion is one of the leading causes for bus delays (H S Levinson 1991; Guerts, Schaufeli, and Buunk 1993; Cham 2006; Xuan, Argote, and Daganzo 2011).

For those internal factors and external factors, some researchers have investigated the effects of different service improvement strategies on service reliability. El-Geneidy conducted a series of visual and analytical analyses to identify causes of decline in reliability levels. Results showed that schedule revisions were needed to improve run time and schedule adherence, and stop consolidation were needed to decrease variability of service through concentrating passenger demand along a fewer number of stops (A.M. El-Geneidy, Horning, and Krizek 2011). Diab et al. examined the impacts of various improvement strategies, implemented along one heavily utilized bus route, on running time deviation from schedule, variation in running time,

and variation in running time deviation from schedules. They found that: (a) the introduction of a smart card fare collection system increased bus running time and service variation; (b) articulated buses, limited-stop bus service and reserved bus lanes had mixed effects on variation in comparison to the running time changes; and (c) TSP did not show an impact on variations (E. Diab and El-Geneidy 2013).

2.3 Bus Arrival Time Prediction

Providing transit users with reliable and accurate travel information can enhance TSR (Vanajakshi, Subramanian, and Sivanandan 2009). While, bus arrival time information is the most preferred information by transit users. The provision of bus arrival time information to passengers accurately is vital. It will be useful to passengers to reduce the waiting time at bus stops or to make reasonable travel arrangements before making a trip, thus, more people can be attracted to use public transport (Jeong and Rilett 2004). In order to provide bus arrival time accurately and timely, a variety of bus arrival time prediction models have been developed over the years. The most widely used bus arrival time prediction models can be classified into four categories as below.

2.3.1 Historical Data Based Models

The simplest way to predict bus arrival time is the historical data-based average prediction model. It predicts the bus arrival time from the historical bus travel time of previous journeys with an assumption that the current traffic condition remain stationary. So the result of this model can be reliable only when the traffic condition is relatively stable (Jeong and Rilett 2004).

2.3.2 Regression Models

Regression models are one kind of classical statistical analysis. With a linear function formed by a set of independent variables, it usually can be used to predict and correlate a dependent variable. Frechette et al. proposed a Bayesian regression analysis to estimate vehicular travel times between links in CBD locations using data collected by a video camera. Volume of through way, left and right turning vehicles, number of signalized intersections, percentage of stopped vehicles on each link, and the percentage of heavy vehicles were used as independent variables. The results were quite appreciable but the data extraction from the video camera would be a difficult process (Frechette and Khan 1998). The study from Patnaik et al. used number of stops, number of boarding and alighting passengers, distance, dwell times and weather descriptors as independent variables with a multivariate linear regression model to estimate bus arrival time between time-points along a bus route. The data used in this study were obtained from the APC system installed on buses, and they found that the models they used could be applied to estimate bus arrival time at downstream stops (Patnaik, Chien, and Bladihas. 2004). However, variables in transportation systems are highly correlated, so the applicability of regression models is limited (Chien, Ding, and Wei. 2002).

2.3.3 Kalman Filtering Models

Kalman filters have also been used for bus arrival time prediction. Bae et al. proposed a Kalman filter model to estimate arterial travel time for buses using AVL data, and a bus arrival time prediction model was developed based on the dynamics of both single and multiple stops. They also considered time varying passenger boarding and alighting rates as one input for the model (Bae and Kachroo 1995). The study from Wall et al. used both AVL data and historical data to estimate bus arrival time in Seattle. Two components are involved in their algorithm:

tracking component and prediction component. They used a Kalman filter model to track a vehicle location as well as a statistical estimations technique for the prediction component. The result indicated that the bus arrival time prediction rate is around 78% (Wall and Dailey 1999). Also, Shalaby et al. used both AVL data and APC data with the Kalman filtering technique to predict bus travel time, and the result indicated that this model can predict bus arrival times up to an hour in advance (Shalaby and Farhan 2003).

2.3.4 Artificial Neural Network Models

Artificial Neural Networks (ANN) is an effective countermeasure to deal with complex relationships between predictors that can arise within large amounts of data, process non-linear relationships between predictors, and process complex and noise data (Jeong and Rilett 2004). Gurmu et al. only used the GPS data as the input for their dynamic bus arrival time prediction ANN model. The results indicated that this model can provide accurate prediction of bus arrival time as a given downstream bus stop, and compared with other historical bus arrival time prediction models, both the prediction accuracy and robustness were outperformed (Gurmu and Fan 2014).

2.4 Summary of Literature Review

In this chapter, the definition and measures of TSR, the previous transit schedule adherence analysis and the methods of bus arrival time prediction are reviewed.

There are a number of internal and external factors affecting the bus on-time reliability, like passenger demand and its distribution along the route, traffic conditions (recurrent traffic congestion and non-recurrent incidences), driving habits, weather conditions, transit operations as well as network design (Sterman and Schofer 1976; Kjmpel 2001). However, as most

researches only focus on one kind of factors, the comparison among different kinds of factors affecting the on-time performance is missed. Also, most previous researches focused on two categories: on time or not on time, and for this kind of requirement, the binominal logit model is suitable. However, in this study, three discrete categories are used: on time, early, and late, which means there are more than two discrete outcomes, therefore, a multinomial logit model is chosen.

For the bus arrival time prediction problem, a lot of methods can be classified as the empirical analysis, like the historical data based models (Jeong and Rilett 2004), the regression models (Frechette and Khan 1998), the Kalman filtering models (Bae and Kachroo 1995) and the ANN models (Jeong and Rilett 2004). However, the empirical analysis has some shortages. For example, the historical data based models are only reliable only when the traffic pattern in the area of interest is relatively stable. The regression models are reliable only when such equations can be established, which may not be possible for many application environments where many of the system variables are typically correlated. The ANN models are more like a black box, and the mechanism of analysis is hidden, which is hard to express the result. Ideally, the empirical models use the historical data to train the models, and then use the trained models to predict. However, with advances in information technologies in ITS, the availability of public transit data has been increasing in the past decades, which has gradually shifted the public transit system into a data-rich environment. Furthermore, most of them are generated in real time, so bus arrival time prediction problem can be solved by some analytical models with these real time transit data instead of the historical data.

Now, a range of widespread applications of AVL, APC, smart bus and other intelligent transportation systems have been integrated into the ETS, which can provide multi-source data

for transit system monitoring and improvement. However, to date, there is little effort to employ the collected data in evaluating bus on-time performance and bus arrival time prediction in Edmonton. The intention of this study is to further extend these studies by using a large scale multi-source data to assess the bus on-time performance and predict bus arrival time based on the real-time GPS and VDS data.

CHAPTER 3 DATA PREPROCESSING

3.1 Test Site Description

Edmonton is the most northerly city in North America with a metropolitan population of over one million as well as the capital of the Canadian province of Alberta. The ETS is the public transit agency which is operated by the City of Edmonton and provides several kinds of transit services, such as the Light Rail Train (LRT) services, regular buses services and door-to-door Disabled Adult Transit Service (DATS). Figure 1 shows the service day map of the ETS, and as of 2015, the ETS has 209 bus routes and 6803 bus stops. The ETS services an area size of 700 sq. km and area population of 817,498. Service is provided between 5 AM to 2 AM, and incorporates morning and afternoon peak service, off-peak service throughout weekdays, evening and late nights, and service throughout weekend and holiday operating periods. With this kind of multiple transit services, suburban feeders can run to a transit center and then transfer to a base route/LRT to the city center or the university, also some feeder routes provide direct express service to and from the city center.

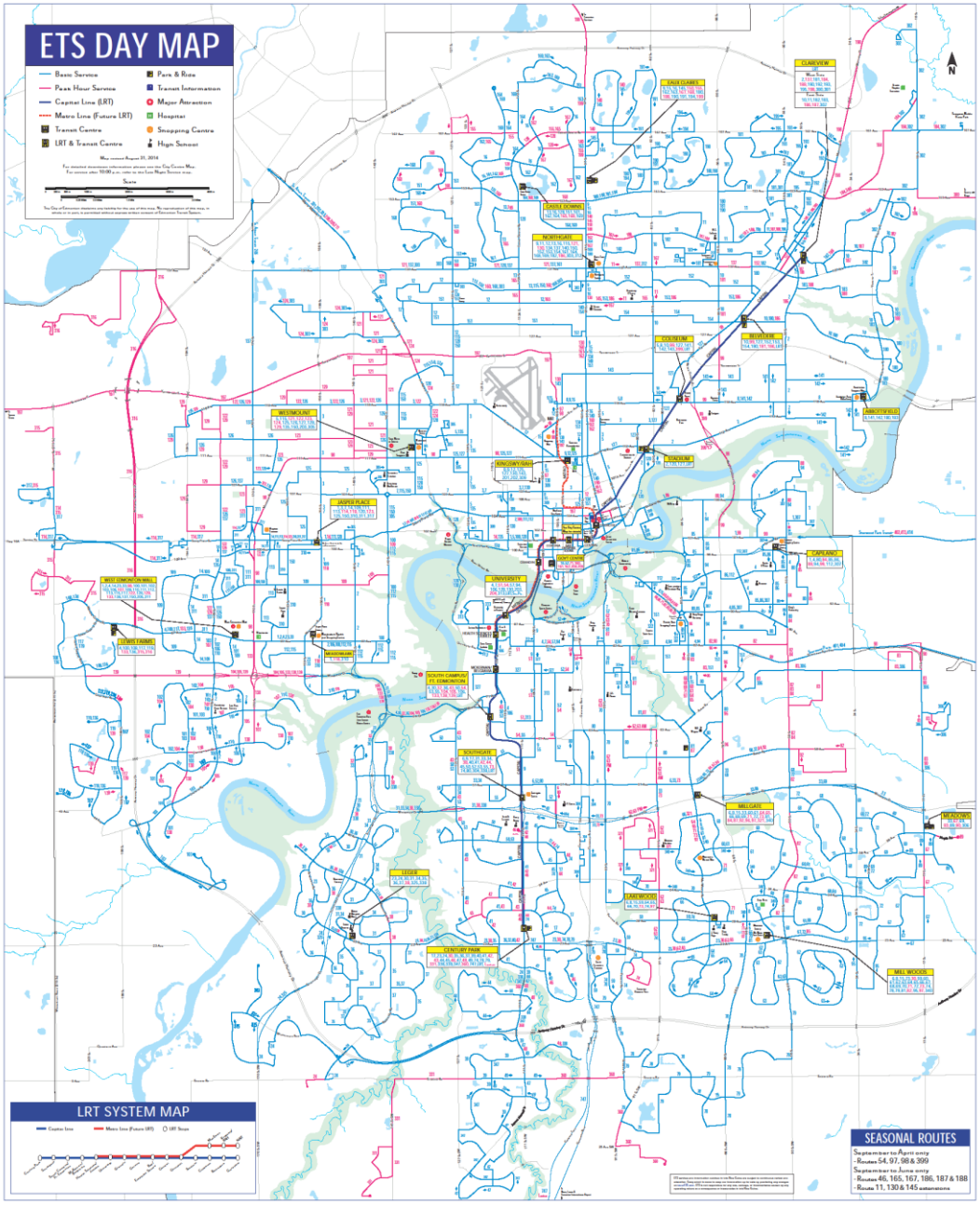

Figure 1 Service Day Map of ETS (City of Edmonton 2015d)

3.2 Data Description

Currently, a range of widespread applications of AVL, APC and other intelligent transportation systems have been integrated in the ETS and five kinds of data will be used in this

study: VDS data, APC data, General Transit Feed Specification (GTFS) data, weather data and road network Geographic Information System (GIS) data.

3.2.1 APC Data

APC System is used by a lot of transportation agencies to collect the number of boarding and alighting passengers for a bus at each stop (Furth et al. 2006). Figure 2 is a snapshot of the APC data management system used in ETS. Reports of link travel time and on time performance by signup were extracted for analyses in this thesis. As shown in Figure 3, the APC system has been implemented in ETS to obtain bus occupancy along with other information such as date and time, calendar event, route number, run number, bus type, bus stop, location type, scheduled time of arrival, recorded time of arrival, recorded time of departure, adherence error and depart load.

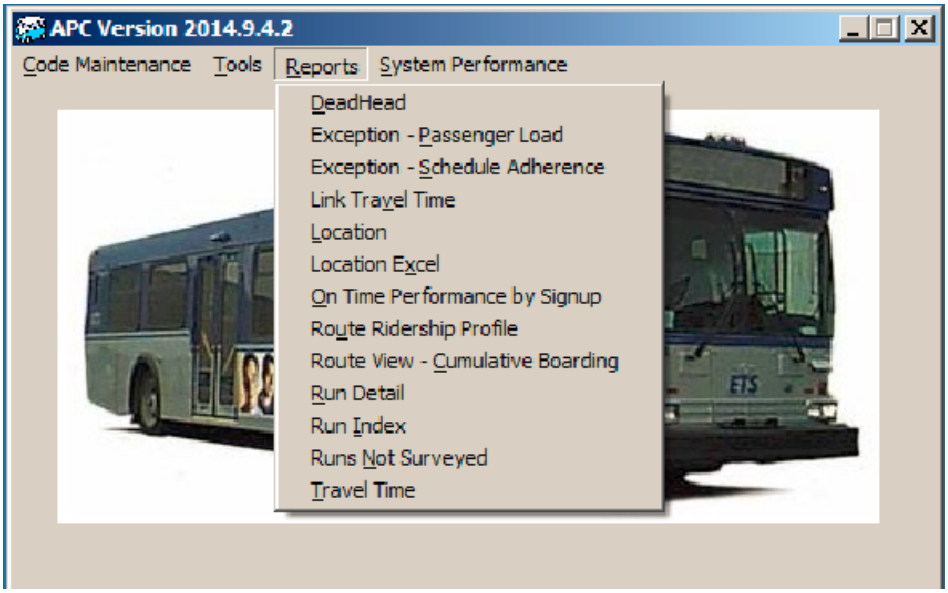

Figure 2 APC System

ETS Automatic Passenger Count
Exception - Schedule Adherence
Signup: SEP13
Start Date: Sep 1 2013
End Date: Nov 30 2013

Service	Survey Date	Calendar Event	Route	Run	Bus Type	Time Period	Bus Stop	Location	Type	Pos	Sched Time	Arrival Time	Depart Time	Time Variance	Adherence Error	Depart Load	Signup	SL Sort	Day Of Week	DOW Sort	TP Sort	Min Value	Max Value	Seating Capacity	RS Sort
Weekday	2013-09-03	E	1	102	Regular	Midday (09:00-14:59)	1780	103 Street/102 Avenue	D	SF	11:13	11:09:03	11:09:58	-3	Time variance smaller than minimum value	13	SEP13	1	TUE	3	3	-2	5	37	361
Weekday	2013-09-03	E	1	102	Regular	Midday (09:00-14:59)	5101	Jasper Place Transit Centre	AN	SF	11:36	11:30:23	11:35:00	-6	Time variance smaller than minimum value	10	SEP13	1	TUE	3	3	-5	10	37	389
Weekday	2013-09-03	E	1	102	Regular	Midday (09:00-14:59)	5101	Jasper Place Transit Centre	DN	SF	11:38	11:30:23	11:35:00	-3	Time variance smaller than minimum value	10	SEP13	1	TUE	3	3	-2	5	37	390
Weekday	2013-09-03	E	1	102	Regular	Midday (09:00-14:59)	5301	Meadowlark Transit Centre	D	DN	11:44	11:40:23	11:40:41	-3	Time variance smaller than minimum value	12	SEP13	1	TUE	3	3	-2	5	37	398
Weekday	2013-09-03	E	1	102	Regular	Midday (09:00-14:59)	5110	Jasper Place Transit Centre	DN	NF	12:15	12:07:52	12:12:03	-3	Time variance smaller than minimum value	10	SEP13	1	TUE	3	3	-2	5	37	421
Weekday	2013-09-03	E	1	102	Regular	Midday (09:00-14:59)	1242	124 Street/102 Avenue	D	SF	12:24	12:20:07	12:20:15	-4	Time variance smaller than minimum value	11	SEP13	1	TUE	3	3	-2	5	37	437
Weekday	2013-09-03	E	1	102	Regular	Midday (09:00-14:59)	1746	122 Street/102 Avenue	D	WF	13:54	13:59:40	14:00:03	6	Time variance greater than maximum value	62	SEP13	1	TUE	3	3	-2	5	37	534
Weekday	2013-09-03	E	1	102	Regular	Midday (09:00-14:59)	5101	Jasper Place Transit Centre	DN	SF	14:08	14:13:11	14:14:15	6	Time variance greater than maximum value	50	SEP13	1	TUE	3	3	-2	5	37	552
Weekday	2013-09-03	E	1	102	Regular	Midday (09:00-14:59)	5301	Meadowlark Transit Centre	D	DN	14:14	14:20:48	14:21:09	7	Time variance greater than maximum value	41	SEP13	1	TUE	3	3	-2	5	37	560
Weekday	2013-09-03	E	1	102	Regular	Early Evening (18:00-21:59)	2301	Capilano Transit Centre	AN	WN	18:39	18:51:40	18:52:26	13	Time variance greater than maximum value	13	SEP13	1	TUE	3	5	-5	10	37	811
Weekday	2013-09-03	E	1	106	Regular	AM Peak (05:30-08:59)	2301	Capilano Transit Centre	DN	WN	06:36	06:20:32	06:26:31	-9	Time variance smaller than minimum value	0	SEP13	1	TUE	3	2	-2	10	37	84
Weekday	2013-09-03	E	1	106	Regular	AM Peak (05:30-08:59)	1780	103 Street/102 Avenue	D	SF	07:12	07:04:58	07:07:14	-5	Time variance smaller than minimum value	16	SEP13	1	TUE	3	2	-2	10	37	119
Weekday	2013-09-03	E	1	106	Regular	Midday (09:00-14:59)	2301	Capilano Transit Centre	DN	WN	09:13	09:04:01	09:09:01	-4	Time variance smaller than minimum value	9	SEP13	1	TUE	3	3	-2	5	37	246
Weekday	2013-09-03	E	1	106	Regular	Midday (09:00-14:59)	2301	Capilano Transit Centre	DN	WN	11:43	11:33:27	11:38:26	-5	Time variance smaller than minimum value	15	SEP13	1	TUE	3	3	-2	5	37	408
Weekday	2013-09-03	E	1	107	Regular	AM Peak (05:30-08:59)	2301	Capilano Transit Centre	AN	WN	08:23	08:41:43	08:42:27	19	Time variance greater than maximum value	20	SEP13	1	TUE	3	2	-5	10	37	163
Weekday	2013-09-03	E	1	107	Regular	Owl (25:00-31:59)	5101	Jasper Place Transit Centre	AN	SF	25:21	25:15:03	25:28:42	-6	Time variance smaller than minimum value	25	SEP13	1	TUE	3	7	-5	10	37	1197
Weekday	2013-09-03	E	1	117	Regular	PM Peak (15:00-17:59)	5009	West Edmonton Mall Transit Centre	DN	EF	15:40	15:33:46	15:33:46	-6	Time variance smaller than minimum value	12	SEP13	1	TUE	3	4	-2	10	37	82
Weekday	2013-09-03	E	1	117	Regular	Early Evening (18:00-21:59)	5101	Jasper Place Transit Centre	DN	SF	18:08	18:05:14	18:05:14	-3	Time variance smaller than minimum value	37	SEP13	1	TUE	3	5	-2	5	37	228

Figure 3 Samples of APC Data

3.2.2 Weather Data

The weather data is achieved from Environment Canada (Environment Canada 2015). It provides an online access to the historical hourly and daily archived weather data at various weather stations across Canada. As shown in Figure 4 and Figure 5, typical observations made at each station include air temperature, dew point temperature, relative humidity, precipitation type, visibility, and wind speed, all on an hourly basis with the exception of the precipitation intensity and snow on ground, which are available in daily totals. The reported hourly weather phenomena in winter include snow, snow grains, clear, cloudy, ice crystals, ice pellets, ice pellet showers, snow showers, snow pellets, fog, ice fog, blowing snow, freezing fog, and other.

Date/Time	Temp (°C)	Dew Point Temp (°C)	Rel Hum (%)	Wind Dir (10s deg)	Wind Spd (km/h)	Visibility (km)	Stn Press (kPa)	Wind Chill	Weather
1/1/2013 0:00	0	-5.2	68	27	18	24.1	93	-5	NA
1/1/2013 1:00	-1.4	-6.2	70	29	16	24.1	93.04	-6	NA
1/1/2013 2:00	-0.6	-7	62	28	18	24.1	93.09	-6	Mostly Cloudy
1/1/2013 3:00	-1	-7.8	60	28	23	24.1	93.16	-7	NA
1/1/2013 4:00	-0.8	-7.4	61	29	26	24.1	93.23	-7	NA
1/1/2013 5:00	-0.3	-6.1	65	29	26	24.1	93.26	-6	Snow Showers
1/1/2013 6:00	0.3	-5.3	66	29	32	24.1	93.27		Snow Showers
1/1/2013 7:00	0.2	-5.4	66	29	21	24.1	93.32		Snow Showers
1/1/2013 8:00	-0.6	-6.6	64	28	21	24.1	93.31	-6	Mostly Cloudy
1/1/2013 9:00	-0.7	-7.1	62	27	22	24.1	93.35	-6	NA
1/1/2013 10:00	-2.5	-8.4	64	27	27	24.1	93.43	-9	NA
1/1/2013 11:00	-2	-8.8	60	27	19	24.1	93.46	-8	Mainly Clear
1/1/2013 12:00	-0.8	-8.3	57	28	25	24.1	93.45	-7	NA
1/1/2013 13:00	-1.1	-8.1	59	29	21	24.1	93.49	-7	NA

Figure 4 Samples of Hourly Weather Data

Date/Time	Max Temp (°C)	Min Temp (°C)	Mean Temp (°C)	Heat Deg Days (°C)	Total Rain (mm)	Total Snow (cm)	Total Precip (mm)	Snow on Grnd (cm)
1/1/2013	0.8	-7.2	-3.2	21.2	0	0	0	21
1/2/2013	-1.4	-16.9	-9.2	27.2	0	0	0	21
1/3/2013	0.3	-12.9	-6.3	24.3	0	0	0	21
1/4/2013	-6	-16.9	-11.5	29.5	0	0	0	21
1/5/2013	-2.9	-13.3	-8.1	26.1	0	0	0	21
1/6/2013	-0.8	-10.5	-5.7	23.7	0	0.4	0.4	21
1/7/2013	-3.5	-14.2	-8.9	26.9	0	0	0	22
1/8/2013	-0.3	-12.3	-6.3	24.3	0	0	0	22
1/9/2013	-3.8	-7.2	-5.5	23.5	0	0	0	21
1/10/2013	-7.2	-25.2	-16.2	34.2	0	0.4	0.2	22
1/11/2013	-16.7	-25.9	-21.3	39.3	0	0	0	22
1/12/2013	-9	-24.9	-17	35	0	1.4	1	22
1/13/2013	-9.2	-22.5	-15.9	33.9	0	1	0.4	23
1/14/2013	3.7	-12	-4.2	22.2	0.2	5.2	3.4	24
1/15/2013	8.3	1.5	4.9	13.1	0	0	0	28
1/16/2013	6.4	-3.7	1.4	16.6	0	0.4	0.2	19

Figure 5 Samples of Daily Weather Data

3.2.3 GTFS Data

GTFS is short for General Transit Feed Specification, which is a format definition for public transportation schedules and corresponding geographic information (Google Inc. 2015b). There are two kinds of GTFS feeds: static and dynamic. The static GTFS feeds include the transit information like trips, stops, routes and other schedule transit data. Currently, the City of Edmonton uses three kinds of static GTFS feeds: bus stop GTFS feed, bus schedule GTFS feed and bus trip GTFS feed. As shown in Figure 6, the bus stop GTFS feed includes bus stop ID, bus stop name, latitude and longitude. Figure 7 is the samples of bus route GTFS feed, and the bus route GTFS feed includes bus route ID and bus route name. As shown in Figure 8, the bus trip GTFS feed has bus route ID, bus service ID and bus trip ID.

stop_id	stop_name	stop_lat	stop_lon
1001	Abbottsfield Transit Centre	53.571965	-113.390362
1002	Abbottsfield Transit Centre	53.572153	-113.3899
1003	Abbottsfield Transit Centre	53.572042	-113.389775
1010	50 Street & 122 Avenue	53.577097	-113.418289
1014	89 Street & 118 Avenue	53.570532	-113.47826
1015	Wayne Gretzky Dr S & 118 Avenue	53.570569	-113.453962
1017	73 Street & 111 Avenue	53.56155	-113.448412
1019	101 Street & 104 Avenue	53.546627	-113.493568
1020	38 Street & 114 Avenue	53.566525	-113.400047

Figure 6 Samples of Bus Stop GTFS Feed (City of Edmonton 2015b)

route_id	route_short_name	route_long_name
140	140	Northgate - Downtown - Lago Lindo
149	149	Eaux Claires- Clareview Late night
122	122	Westmount - WEM
14	14	Jasper Place - WEM
192	192	Brintnell - West Clareview
773	773	Mill Woods/Greenview - Wagner
33	33	Meadows - Mgte - Sgte - Riverbend - WEM
862	862	Ottewell - Austin O'Brien - Burnewood
323	323	Bonnie Doon - Good Sam 96 St
72	72	Millgate - Silver Berry - Mill Woods TC

Figure 7 Samples of Bus Route GTFS Feed (City of Edmonton 2015a)

route_id	service_id	trip_id
137	137-Weekday-2-DEC15-1111100	10548310
69	69-Sunday-2-DEC15-0000001	10612332
134	134-Weekday-3-DEC15-1111100	10643216
3	3-Weekday-3-DEC15-1111100	10700886
190	190-Sunday-1-SEP15-0000001	10087847
121	121-Weekday-2-DEC15-1111100	10545407
180	180-Sunday-1-DEC15-0000001	10479518

Figure 8 Samples of Bus Trip GTFS Feed (City of Edmonton 2015c)

The dynamic GTFS feed is also called GTFS-realtime feed. GTFS-realtime is an extension to GTFS, and it is used for many transportation agencies as a feed specification to update transit's real-time information. The GTFS-realtime specification categorizes three types of transit real-time information: (1) trip updates, (2) service alerts and (3) vehicle positions

(Google Inc. 2015a). The smart bus GPS data is stored in the data structure of vehicle positions, and the GPS location for each smart bus will be recorded per cycle period. Also, the trip ID, the timestamp and the bus label will be stored. Currently, the City of Edmonton has equipped smart bus technology on 25 bus routes as of October 2015, and this GTFS-realtime standard has been used to publish the smart bus real time data since July 2015 (City of Edmonton 2015g). As shown in Figure 9, two kinds of data are published now: (1) Real time vehicle position data and (2) Real time trip update data. The real time vehicle position data stores the bus's real time position data per 30 seconds, like the trip ID, vehicle ID, latitude, longitude, speed and timestamp. For the real time trip update data, it records the bus's real time departure time and arrival time for each bus stop.

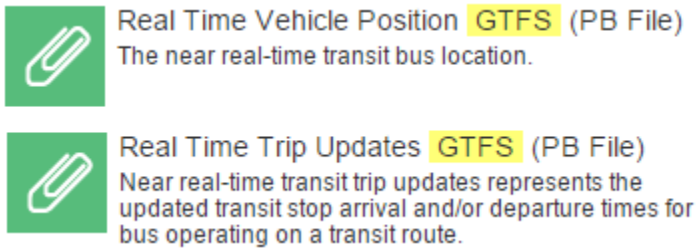

Figure 9 The Real Time Bus Data of Edmonton (City of Edmonton 2015e)

3.2.4 VDS Data

VDS use loop detectors installed on the road to collect the traffic data. The collected data includes the traffic volume, traffic occupancy and traffic speed. As shown in Figure 10, for each lane, the corresponding loop detector detects the data including timestamp, loop detector ID, lane ID, volume, occupancy and speed per cycle period.

recordId	datetime	vdsId	lane	volume	occupancy	speed
394998703	2014-02-06 15:30:20.000	1004	1	540	4	76
394998704	2014-02-06 15:30:20.000	1004	2	1350	10	91
394998705	2014-02-06 15:30:20.000	1004	3	1350	12	86
394998706	2014-02-06 15:30:20.000	1003	1	900	55	-1
394998707	2014-02-06 15:30:20.000	1003	2	540	4	82
394998708	2014-02-06 15:30:20.000	1003	3	0	0	0
394998709	2014-02-06 15:30:20.000	1002	1	540	3	87
394998710	2014-02-06 15:30:20.000	1002	2	180	1	100
394998711	2014-02-06 15:30:20.000	1002	3	-180	0	81
394998712	2014-02-06 15:30:20.000	1001	1	360	2	72

Figure 10 Samples of VDS Data

As shown in Figure 11, there are 44 loop detectors installed in the citywide now. Each loop detector records the traffic information per 20 seconds, including the volume, speed, occupancy, loop detector ID and timestamp as showed in Figure 10.

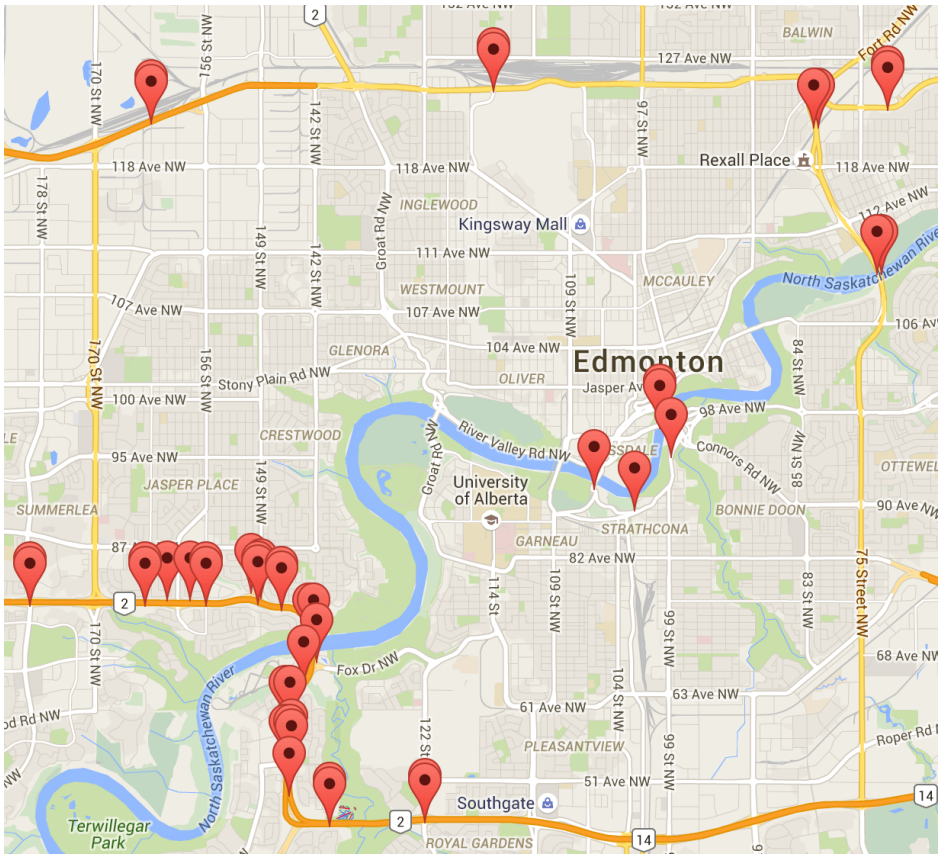

Figure 11 Loop Detector Locations in Edmonton

3.2.5 Road Network Data

The last dataset is the road network GIS shape file, which records the entire road segment's location for the whole road network. As shown in Figure 12, the GIS information stored in the road network shape file contains road segment ID, street name, start intersecting road name, end intersecting road name, street direction, road segment length and road coordinate. Using the road network as the base map, the VDS data, APC data and smart bus GPS data can be mapped on the map.

ID	318686
LRS_DATUM_	25548.00000
STREET_NAM	WHITEMUD DRIVE NW
CARRIAGEWA	WB-S
ROAD_SEGME	NORMAL
SPECIAL_KE	N/A
SEQUENCE_N	36.00000
AUTO_SEQUE	36.00000
START_MEAS	18611.98112
END_MEASUR	19176.90855
GAP_MEASUR	0.00000
START_MARK	39942.00000
START_MAR1	149 STREET NW [Southbound] {Entrance Ramp}
START_MAR2	Y
END_MARKER	124302.00000
END_MARKE1	159 STREET NW {Exit Ramp}
END_MARKE2	Y
ROAD_SEGM1	148605.00000
ROAD_SEGM2	22363.00000
DIGITIZING	2F
START_NODE	16755.00000
END_NODE	16641.00000
FUNCTIONAL	A-A
RESPONSIBL	COE
TEMPORAL_S	CURRENT
ADDRESS_LE	NULL
ADDRESS_L1	NULL
ODD_EVEN_L	ODD
ADDRESS_RI	NULL
ADDRESS_R1	NULL
ODD_EVEN_R	EVEN
LATERAL_OF	21.00000
ROAD_SEGM3	46491.00000
ROAD_SEGM4	22051.00000
X_OR_Y_SOR	X
START_X_CO	27613.10965
START_Y_CO	5931138.20154
END_X_COOR	27052.52584
END_Y_COOR	5931092.57415
LRS_CORREC	NULL
LRS_CORRE1	M
GEOMETRY_L	564.92744
EFFECTIVE_	2012-04-10
EFFECTIVE1	NULL
LTT_FEATUR	25461
LTT_STATUS	NULL
LTT_CURREN	0
LTT_UNIQUE	15599730.00000

Figure 12 Samples of Road Network GIS Data

3.3 Data Integration

An integrated data is critical for both evaluation of factors affecting bus on-time performance and bus arrival time prediction. As the evaluation of factors affecting bus on-time performance is based on bus stop level and bus arrival time prediction focuses on route level, the corresponding integrated data should be extracted by different location condition and time condition as well as the analysis result can be visualized on the road network. The process of data integration is divided into three parts: (a) Spatial data representation, (b) Uniform Location Reference System (LRS) establishment, (c) Topology relationship establishment.

3.3.1 Spatial Data Representation

The first procedure is the spatial data representation process, which is the foundation in location storage. Before a location can be used to tackle real world transportation problems, data must be properly represented in a uniform spatial schema. So the data representation task is going to map each transportation element (including loop detectors, bus stops, smart bus GPS points and road network) into a uniform schema of GIS identification. Two kinds of abstraction type are defined for this uniform schema of GIS identification:

- Point transportation element
- Polyline transportation element

Following this classification rule, each loop detector, bus stop, and smart bus GPS point are classified into the type of point transportation element. For the road network, because it is the base spatial map for the other data, both the road segments and the intersections in the road network need to be classified as well, and each road segment is abstracted as a polyline transportation element as well as each intersection is abstracted as a point transportation element.

3.3.2 Uniform LRS Establishment

The second process is to establish a uniform LRS for those data. LRS existed long before GIS and computers and it is defined as a system of determining the position of an entity relative to other entities to some external frame of reference. With the LRS data model, location meaning can be reflected. There are two kinds of LRS models, one is each transportation element refers to a reference object (an intersection or a road segment) in the road network, and the other one is that each kind of transportation element use their individual geographic coordinate system. In this study, the second kind of LRS data model is chosen, and the reason is that the location information of loop detector data, GTFS data, smart bus GPS data and road network GIS data are composed of a pair of latitude and longitude.

3.3.3 Topology Relationship Establishment

The third process of data integration is the topology relationships establishment. As the evaluation of factors affecting bus on-time performance is focus on stop level and bus arrival time prediction focuses on route level, the integrated data extraction should depend upon network connectivity. The basic elements of GIS points and lines can be thought of as having topology, or defined geometric locations and relationships that represent whether and how the data are connected, so the topology and connectivity of linear features provide the mechanism for representing transportation networks and performing related analysis functions. As the road network is the base spatial map for the other data, the topology relationships establishment is actually going to map each fixed location data (loop detectors, bus stops) and moving location data (bus real time GPS point) on the road network. In this study, the topology relationships can be divided into four sub-topologies:

- Tie each bus stop to a road segment.

- Tie each loop detector to a road segment.
- Tie each smart bus GPS point to a road segment.
- Tie each intersection to its connected road segment.

The mechanism in setting topology relationships is distance calculation. While after the procedure one and procedure two, all the loop detector points, bus stops, smart bus GPS points, intersection points and road segments have been mapped in a uniform schema with their corresponding 2-D coordinates. Therefore, the distance calculation among them can be realized easily.

For the topology between bus stops and road segments, the processing steps for each bus stop are:

1. For all the road segments in the road network, calculate all the distances between this bus stop and them.
2. Order the road segments by the calculated distance from small to large. And choose the smallest one as the road segment for this bus stop.

And the processing to tie each loop detector to a road segment is same as above.

For the topology between intersections and road segments, each road segment is connected with two intersections by direction, one is the start intersection, and the other is the end intersection. However, each intersection can be either start intersection or end intersection for different road segments. As the point coordinate of each intersection is included as either the first point coordinate or the last point coordinate of the polyline coordinate of different road segments, the processing steps to establish topologies between one intersection and its corresponding road segments are:

1. For all the road segments in the road network, calculate all the distances between this intersection and them.
2. After step 1, filter the road segments with calculated distance value of 0.
3. For each road segment in step 2, if the ‘main road name’ and ‘start intersecting road name’ of that road segment is same as the ‘intersection name’ of this intersection, then this intersection is the start intersection of that road segment. Otherwise, this intersection is the end intersection of that road segment.

For each bus route, it has a sequence of bus stops between the start bus stop to the end bus stop, and each bus stop has been mapped on a road segment described above, so each bus route can be connected with a sequence of road segments. The processing to tie each bus GPS point to a road segment is same as the processing to tie each loop detector to a road except that only the road segments included in its route are needed.

Therefore, as shown in Figure 13, after these three procedures above, each loop detector, bus stop, smart bus GPS point, bus route, intersection and road segment have been abstracted into their corresponding GIS object with a location coordinate in a uniform LRS, and the topology among them is set as well. As the weather data is city wide, the connection between weather data and other data is through timestamp.

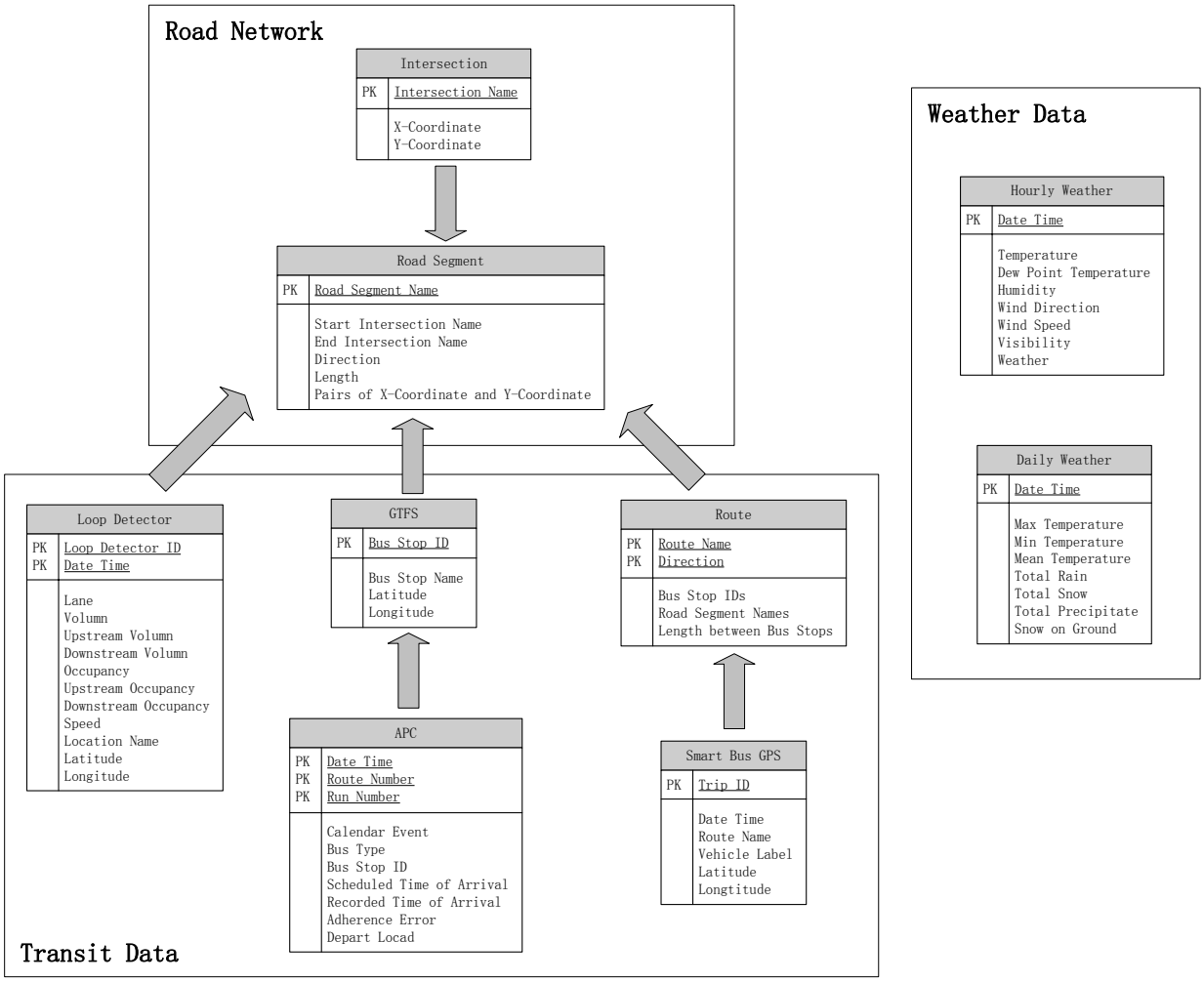

Figure 13 Entity-Relationship of Data Integration

CHAPTER 4 EVALUATION OF FACTORS AFFECTING BUS ON-TIME PERFORMANCE

4.1 Introduction

An understanding of bus service reliability is necessary to develop strategies that help transit agencies provide better services. Edmonton is the most northerly city in North America with a metropolitan population of over one million, where winter lasts from November to March, and varies greatly in length and severity. Adverse weather conditions have an impact on frequency, travel time and headway regularity. However, to date there is little effort to employ the multi-source data in evaluating bus on-time performance in Edmonton. Using the ETS as a case study, this chapter aims to investigate the use of multiple data sources to quantify and determine temporal and weather related bus on-time performance in Edmonton.

APC data, weather data and GTFS data are used for the input of this analysis. With the data integration, each APC data record can be extracted with the weather and GTFS data, and about 0.5 million valid records covering the period from 2011 to 2014 are extracted. For each extracted integrated record, it contains the following information: date and time, calendar event, route number, run number, bus type, bus stop, location type, scheduled time of arrival, recorded time of arrival, recorded time of departure, adherence error, depart load, air temperature, dew point temperature, relative humidity, precipitation type, visibility, wind speed, precipitation intensity, snow on ground, weather condition, X-coordinate, and Y-coordinate.

4.2 Methodology

In this study, a multinomial logit model is conducted to quantify risk factors influencing bus on-time performance.

Three discrete categories are used in this study: on time, early, and late. Since the three possible discrete outcomes have a natural ordering, an ordered probability model is most convenient in this study. Suppose that P_{ni} and U_{ni} refer to the probability of bus n being of schedule adherence i and a linear function that determines the on-time performance, respectively. The probability of the nth bus having a schedule adherence level i is given by

$$P_{ni} = P(U_{ni} \geq U_{ni'}),\ \forall i' \in I, i' \neq i \quad (1)$$

where I denotes a set of all possible, mutually exclusive levels. In this context, the U_{ni} function can be defined as

$$U_{ni} = \beta_i x_n + \varepsilon_{ni} \quad (2)$$

where β_i denotes a vector of estimable parameters, x_n denotes a vector of the observable characteristics that determines severity and ε_{ni} denotes an unobservable random error. If ε_{ni} assumed to have generalized extreme value distribution, the multinomial logit model can be derived as follows:

$$P_{ni} = \frac{e^{\beta_i x_n}}{\sum_{\forall i' \in I} e^{\beta_{i'} x_n}} \quad (3)$$

Given Eq. (3), the coefficient values in vector β can be estimated using standard maximum likelihood methods. Then, the I−1 log-odd ratios of the outcomes become

$$\ln\left(\frac{P_{ni}}{P_{nl}}\right) = \beta_i x_n - \beta_l x_n = (\beta_i - \beta_l) x_n, i = 1, \ldots, I-1 \quad (3)$$

The change in probability in each factor should also be computed to properly examine marginal effects. In this study, the independent variables are coded as 0 and 1 indicator values. Therefore, the probability relative to any of the observed variables cannot be differentiated to

compute a standard elasticity. In this sense, the direct pseudo-elasticity of the probability, namely the percentage change in probability when an indicator variable is changed from 0 to 1, can be determined as

$$E_{x_{nk}}^{P_{ni}} = \frac{P_{ni}[given\ x_{nk}=1] - P_{ni}[given\ x_{nk}=0]}{P_{ni}[given\ x_{nk}=0]} \quad (4)$$

where the kth indicator variable for a bus n, x_{nk} is shifted. The direct pseudo-elasticity for the multinomial model can be briefly defined by inserting Eq. (3) into Eq. (4).

$$E_{x_{nk}}^{P_{ni}} = \left(e^{\beta_{ik}} \frac{\sum e^{\beta_{i'} x_n}}{\sum e^{\Delta(\beta_{i'} x_n)}} - 1 \right) \times 100 \quad (5)$$

where I is the set of possible outcomes, $\Delta(\beta_{i'} x_n)$ is the value with x_{nk} set to 1 and $\beta_{i'} x_n$ is the value with x_{nk} to 0.

The likelihood ratio test can be applied to test if schedule adherence model is significantly different among potential risk factors. The test statistic is given by

$$X^2 = -2[LL(\beta_T) - \sum_G LL(\beta_g)] \quad (6)$$

where $LL(\beta_T)$ is the model's log likelihood at the convergence of the model estimated on all risk factors being tested, $LL(\beta_g)$ is the log likelihood at the convergence of the model estimated on the subset schedule adherence group g and G is the set of all groups. This likelihood ratio test statistic is χ^2 distributed with degrees of freedom equal to the summation of coefficients estimated in the subset data models less the number of coefficients estimated in the total data model.

4.3 Results and Discussions

4.3.1 Temporal Analysis

Buses are considered to be on time when they arrive no more than three minutes earlier or five minutes later than the scheduled arrival times at each bus stop. The temporal variation of bus on-time performance is evaluated by calculating the percentage of on-time arrivals (POA) and the standard deviation of arrival time variance (SDV). As a first step to investigate the temporal variation of bus on-time performance, the POA and SDV for each hourly period for all days of the week are extracted for clear weather conditions. Figure 14 and Figure 15 show the POA and SDV for all hourly intervals and all days of the week under the clear weather condition. The POA during the weekday and PM commuting periods is lower. Periods around midnight on Saturday and Sunday have relatively lower POA compared to weekday nights. This is an expected result considering the higher traffic on weekend nights mainly due to night-out activities. The highest POA is recorded during early morning hours. The lowest POA is observed during PM peak hours and midday on weekdays. Among weekend days, Saturday has lower POA compared to Sunday.

SDV follows more or less the same trend as POA. The higher variations are generally observed in the same days and periods that have lower POA. There is a sharp increase is observed after 8 AM. During PM peaks, the reverse situation is observed. SDV values generally start decreasing after 6 PM whereas POA values sustain high values until 7 PM. These periods can generally be viewed as the traffic congestion build-up and dissipation periods. Hence, SDV reaches its highest value after congestion is fully built up and does not start decreasing until the maximum congestion dissipates.

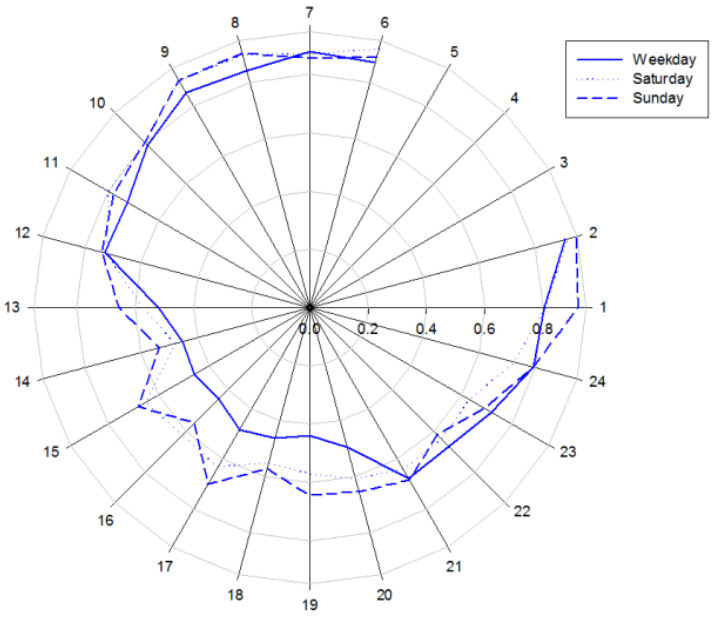

Figure 14 Visualization of POA under Clear Weather

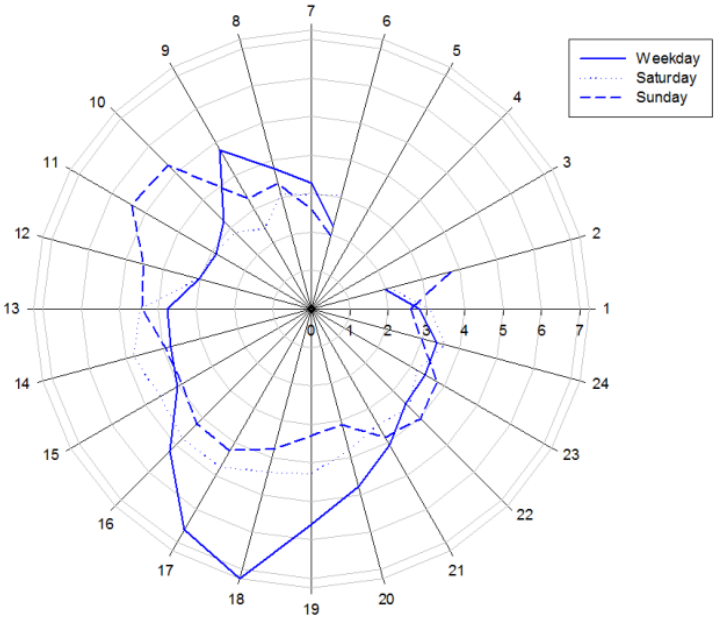

Figure 15 Visualization of SDV under Clear Weather

4.3.2 Impact of Weather

Figure 16 to Figure 21 show the POA and SDV for all hourly intervals under three weather conditions. During the weekday, adverse weather conditions decrease the POA and increase the SDV.

POA decreases for almost all time periods. The change of POA is much higher in the day time and early evening, from 8 AM to 9 PM. The change of SDV is mainly in the positive direction under snow weather condition, which means adverse weather results in lower bus on-time reliability. The change of SDV is much higher during the AM and PM peak periods. The severity of snow conditions affects the magnitude of change in POA and SDV. The decrease of POA under heavy snow condition is a little higher than it under moderate snow condition. However, the decrease of SDV due to the severity of snow conditions has different trend: moderate snow results in higher change in the afternoon and early evening, while heavy snow results in higher change in the morning and noon.

Saturday and Sunday follow more or less the same trend as weekday. First, the snow weather condition decrease the POA, but the magnitude of change is smaller than it in the weekday. Second, the POA does not decrease significantly in the peak periods, compared with the POA change in the peak periods of weekday. Third, the change of SDV has different trends in the weekend, as the change of SDV is sometimes in the negative direction under snow weather conditions.

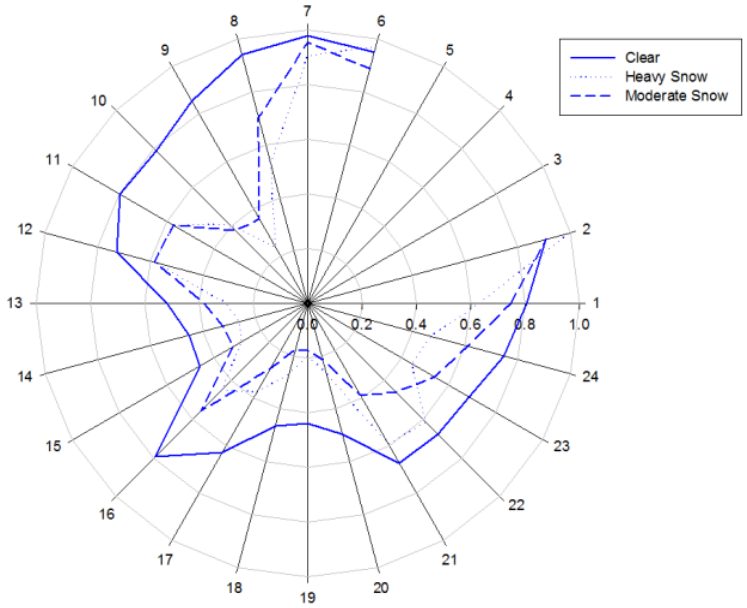

Figure 16 Visualization of POA on Weekday

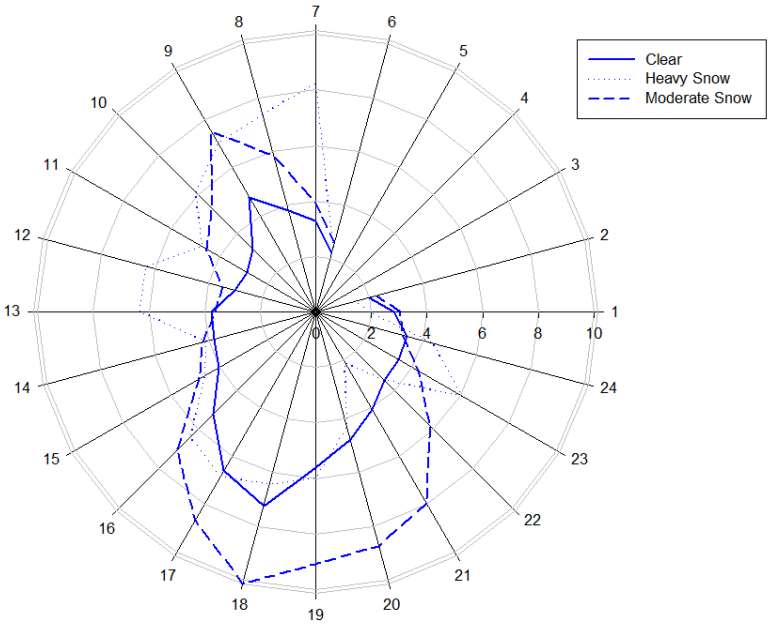

Figure 17 Visualization of SDV on Weekday

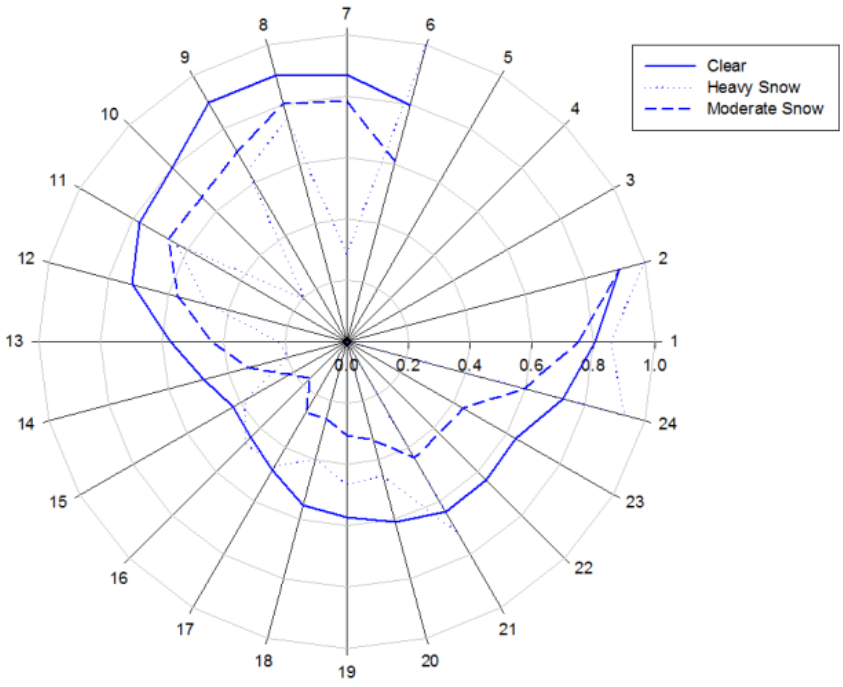

Figure 18 Visualization of POA on Saturday

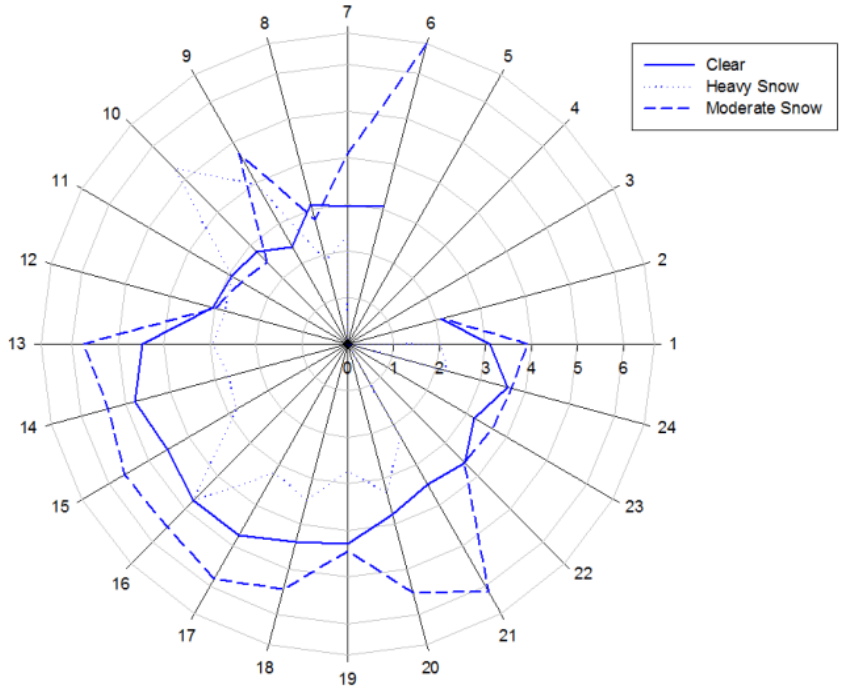

Figure 19 Visualization of SDV on Saturday

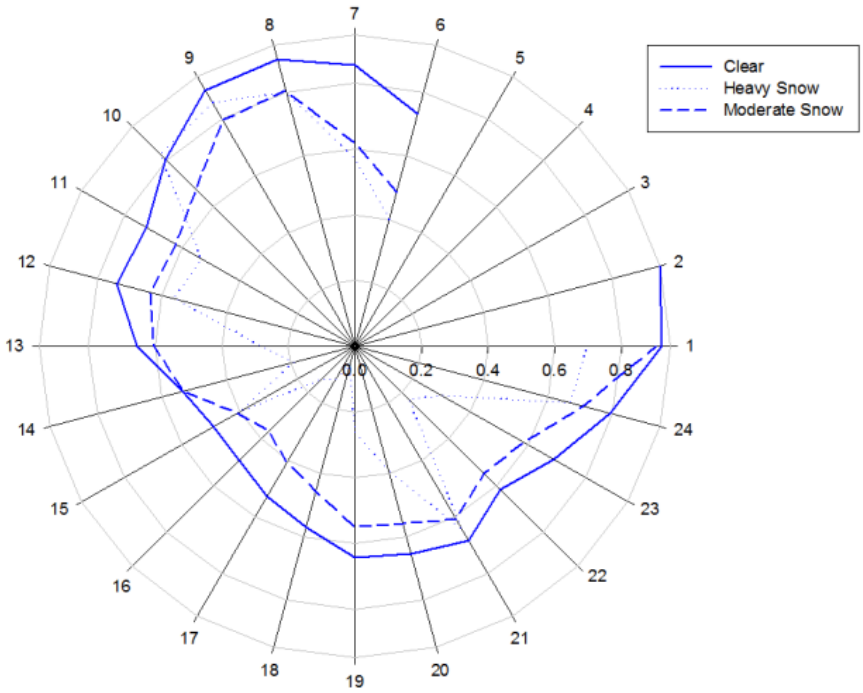

Figure 20 Visualization of POA on Sunday

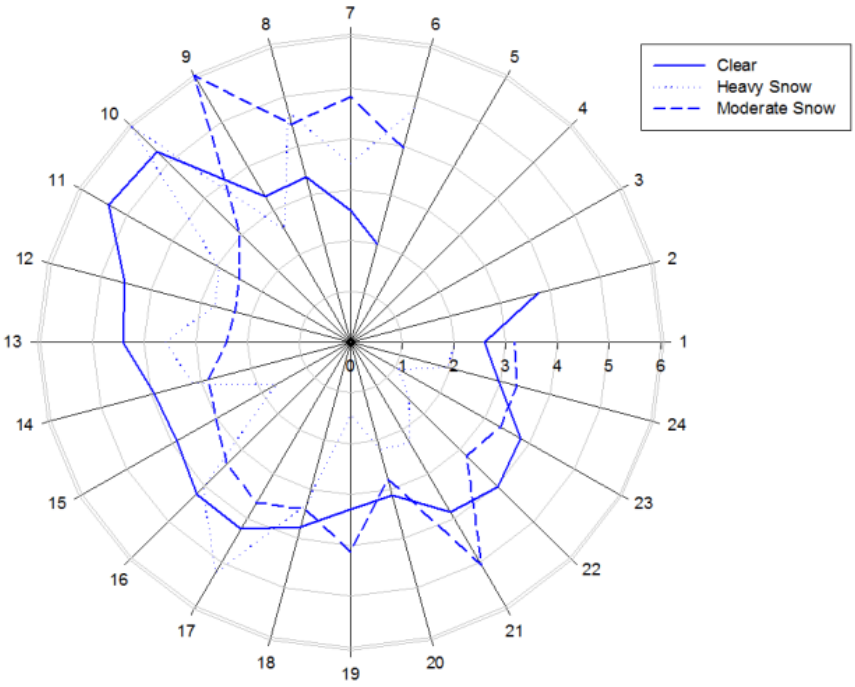

Figure 21 Visualization of SDV on Sunday

4.3.3 Multinomial Logit Analysis

4.3.3.1 On-time Performance Model

The multinomial logit model is implemented to relate the three alternative arrival outcomes to the set of contributing underlying factors. The model is specified on the left-hand side as the log of the relative probabilities of pairs of alternative outcomes. In the present application, with three alternative outcomes, three pair combinations are addressed (on-time/early, on-time/late, and early/late) resulting in three equations. The general specification of the on-time performance model is as follows:

$$\log(\mathrm{P_i}/\mathrm{P_j}) = \mathrm{f}(\mathrm{Count}, \mathrm{Wkdy}, \mathrm{AMP}, \mathrm{PMP}, \mathrm{Msw}, \mathrm{Hsw}, \mathrm{Lvy}, \mathrm{Hdwy}, \mathrm{Rlh}) \quad (7)$$

Where

$\log(P_i/P_j)$ = the log of the relative probabilities of alternative outcomes i and j;

Count = the position of the sampled time point in the sequence of time points on the bus route;

Wkdy = a dummy variable equaling 1 for weekdays, 0 otherwise;

AMP = a dummy variable equaling 1 for AM peak inbound trips, 0 otherwise;

PMP = a dummy variable equaling 1 for PM peak outbound trips, 0 otherwise;

Msw = a dummy variable equaling 1 for moderate snow condition, 0 otherwise;

Hsw = a dummy variable equaling 1 for heavy snow condition, 0 otherwise;

Lvy = a dummy variable equaling 1 for low visibility condition, 0 otherwise;

Hdwy = the scheduled headway (in minutes);

Rlh = the route length (in kilometers)

The position of the time point on the route, represented by the variable "Count," may also be important. Delays or early arrivals in the initial part of the route can contribute to poor on-time performance later in the route, particularly if holding/control actions are not taken.

Given greater variation of internal and external performance-affecting conditions during the week, one would expect greater difficulty in maintaining on-time service on weekdays. This hypothesis is addressed by including a weekday dummy variable. Dummy variables for AM peak (7:00 am-8:59 am) inbound and PM peak (4:00 pm-5:59 pm) outbound trips are also included to assess on-time performance in the situations when ridership levels are highest and traffic and operating conditions pose the greatest challenge to maintaining scheduled service.

As Edmonton suffers from unfavorable winter weather conditions usually from November to March, weather characteristics are considered a prominent factor on bus on-time performance. Three weather characteristics are under consideration: low visibility, moderate snow and heavy snow. Rain conditions are not included in this study due to their low rate of

occurrence in Edmonton, resulting in relatively lower record sample size. Several relevant categories are merged. 80% of all snowy days are considered as heavy, and 20% of all snowy days are considered as moderate.

As previously discussed, both short and long headways may negatively affect on-time performance. We thus include both the linear and quadratic forms of this variable in the model.

4.3.3.2 Results Analysis

The sample observation are 1677 bus arrivals at time points encountered in 200 bus trips on 59 routes in Edmonton. Of the total arrivals, 1370 (81.7%) fell within the on-time range, while 97 observations (5.8%) arrived early and 210 (12.5%) arrived late. The distribution appears to be log normal.

The estimated bus on-time performance model is presented in Table 1. The first two columns of coefficients in the table pertain to the estimated equations for the probabilities of on-time arrival in relation to the alternatives of arriving early and late, respectively. The final column pertains to the equation estimating the relative probabilities of early and late arrivals. Put another way, the first two columns' coefficients test for the previously defined condition establishing the existence of a controllable source of poor on-time performance, while the last column's coefficients test for the directional condition.

Regarding the temporal factors, in the afternoon period the likelihood of both early and late arrivals increased, without a clear tendency toward either type of failure. This outcome was consistent with the finding in Figure 15 that maximum SDV occurred during the PM peak. In the morning period the likelihood of early arrival increased and the likelihood of late arrival decreased, which means on-time performance actually was better during the AM peak period. Day of week was also estimated to affect on-time performance. In the weekend the likelihood of

early arrival increased and the likelihood of late arrival decreased. The on-time performance actually improved during the weekend.

Regarding the weather factors, under the moderate snow condition, the likelihood of late arrivals increased significantly, with a clear tendency toward arrival late failure. The estimates for the visibility variable indicated that late arrivals become more likely with lower visibility.

Regarding the operating factors, a long route resulted in increases in the likelihood of both early and late arrivals. The estimates of stop count indicated that the probability of late arrivals tended to increase as buses progress toward a route's terminal point.

Regarding the overall performance of the on-time model, there is a substantial increase in the log likelihood function over its null value, and the corresponding value of likelihood ratio statistic (2194) easily exceeds the critical χ^2 value of 54 (0.001 level, 26 d.f.). The likelihood ratio index, a loosely interpretable indicator of goodness of fit (whose value ranges from 0 to l), is 0.64.

Table 1 Multinomial Logit Model for Bus On-time Performance (Asymptotic *t* values in parentheses)

Variables	$\log(P_2/P_0)$	$\log(P_2/P_1)$
Intercept	3.35 (4.96)	7.23 (8.37)
AM Peak	-0.30 (-2.62)	1.40 (3.01)
PM Peak	-0.83 (-4.48)	-1.75 (-6.95)
Weekend	-0.43 (-2.91)	0.98 (3.19)
Moderate Snow	1.27 (4.07)	-2.07 (-5.89)
Temperature	-0.02 (-0.195)	-0.09 (-0.77)
Visibility	0.07 (2.36)	-0.31 (-2.92)
Route Length	-0.10 (-2.81)	-0.20 (-2.98)
Stop Count	0.08 (2.66)	-0.20 (-2.82)

Log-likelihood Function (0): - 1705.0 Log-likelihood Function (p): -607.9 Likelihood Ratio Statistic: 2194.2 Likelihood Ratio Index: 0.64 n: 1677
Alternative 0 = early, alternative 1 = late, alternative 2 = on time. Significant at the .001 level.

CHAPTER 5 BUS ARRIVAL TIME PREDICTION

5.1 Introduction

Providing transit users with reliable and accurate travel information can enhance TSR (Vanajakshi, Subramanian, and Sivanandan 2009). While, bus arrival time information is the most preferred information by transit users. The provision of bus arrival time information to passengers accurately is vital. It will be useful to passengers to reduce the waiting time at bus stops or to make reasonable travel arrangements before making a trip, thus, more people can be attracted to use public transport (Jeong and Rilett 2004). As Edmonton has an adverse weather conditions from November to March, an accurate bus arrival time prediction is extremely important.

For the bus arrival time prediction problem, a lot of previous researches focuses on empirical analysis, like the historical data based models, the regression models, the Kalman filtering models and the ANN models. However, the empirical analysis has some shortages. For example, the historical data based models are reliable only when the traffic pattern in the area of interest is relatively stable. The regression models are reliable only when such equations can be established, which may not be possible for many application environments where many of the system variables are typically correlated. The ANN models are more like a black box, and the mechanism of analysis is hidden, which is hard to express the result. Ideally, the empirical models use the historical data to train the models, and then use the trained models to predict. However, with advances in information technologies in ITS, the availability of public transit data has been increasing in the past decades, which has gradually shifted the public transit system into

a data-rich environment, furthermore, most of them are generated in real time, so bus arrival time prediction problem can be solved by some analytical models with these real time transit data instead of the historical data.

Nowadays, a range of transportation agents have installed GPS sensors on the buses to monitor the buses in real time, and one of the universal real time transit data standard provided by the Google is called GTFS-realtime (Google Inc. 2015a). Following this standard, the agents can provide the public lots of transit data in real time. And with the real time transit data and loop detector data, this chapter aims to propose a bus trajectory reconstructive model to predict the bus arrival time based on the integrated multi-source data and evaluate the result.

5.2 Methodology

5.2.1 Real-Time Bus Tracing

In order to reconstruct the trajectory of a moving bus in real time, the vehicle position stored in each GPS record needs to be projected on the transit network and the road network. As shown in Figure 22, the first step is to filter the needed GPS record with an input trip ID, which is recorded in the bus trip ID schedule file. Furthermore, each bus has a unique trip ID for each trip in one day. After the step one, the corresponding latitude and longitude for that vehicle position can be achieved. Due to the GPS location data has some kind of random error, in most times, the achieved latitude and longitude cannot be mapped on the road network correctly. Therefore, the next step is to calibrate the GPS location information onto the route network and road network. As talked in the chapter 3, for each route, its route network has been maintained in a GIS environment with the road network. The bus route is composed of a sequence of road segments, and each pair of latitude and longitude in the GPS location data should be mapped on

one road segment of the route network. The mapping mechanism is same as talked in the chapter 3, and by calculating the distance to all the road segments in the route network, the shortest one is chosen to be the road segment for this GPS point. Therefore, by mapping this GPS point on the road network, the distance from this GPS point to the next bus stop and the distance from this GPS point to the next downstream loop detector can be determined.

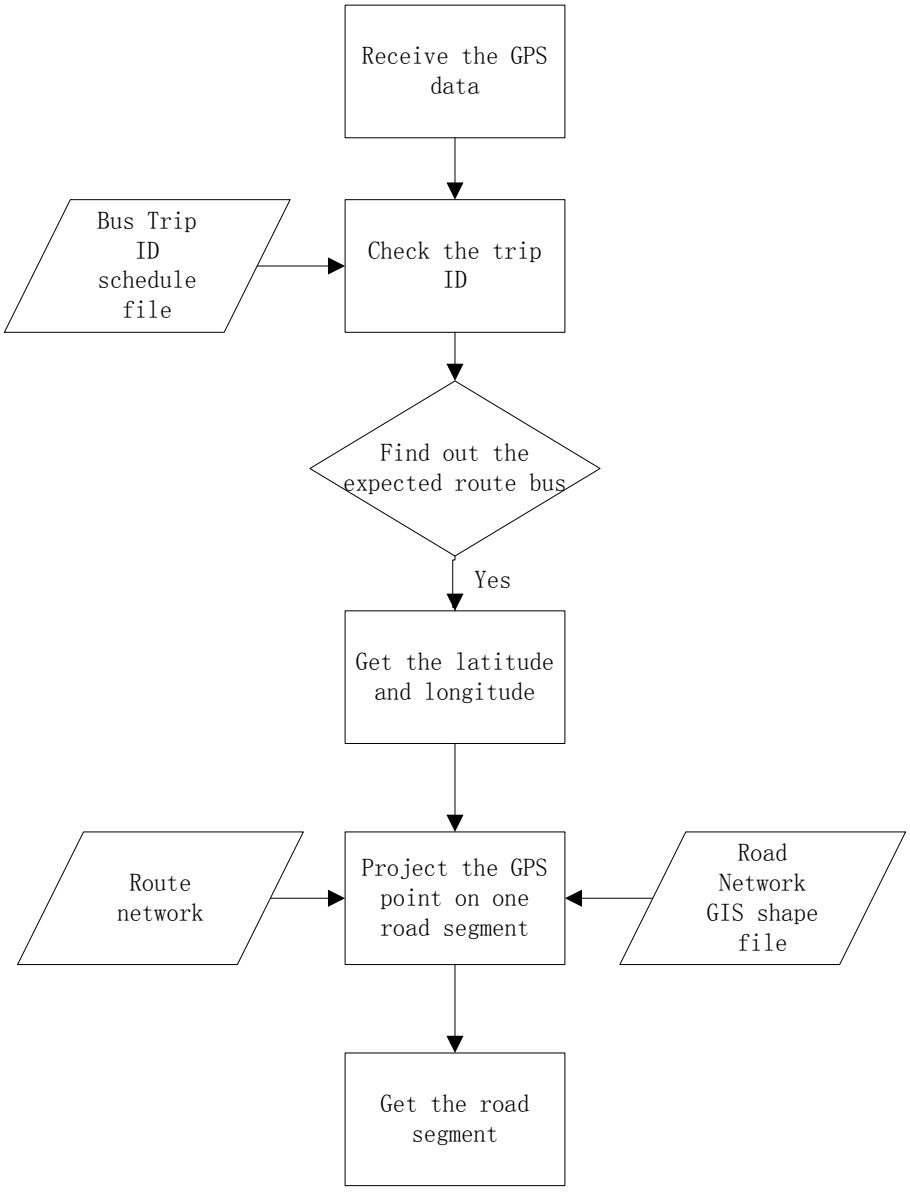

Figure 22 The Workflow of Real-Time Bus Tracing

Using the real-time bus tracing method, each real-time bus GPS point can be visualized as shown in Figure 23.

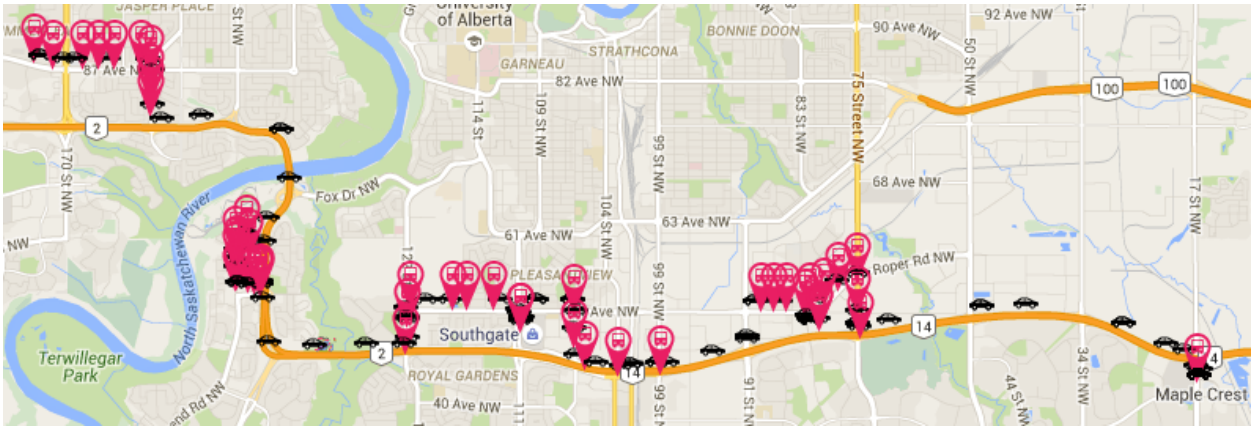

Figure 23 Result Sample of Real-Time Bus Tracing

5.2.2 Real-Time Bus Trajectory Reconstruction

A vehicle trajectory can be thought as a sequence of adjacent sections and by calculating the time and distance of each section, then the whole trajectory can be found. A study from Van et al. forms an algorithm to estimate section level travel time by using the dual loop detectors. In their algorithm, each section is defined as a road segment between the upstream loop detector and the downstream loop detector, and the speed is linearly changed in the section from the speed recorded in the upstream loop detector to the speed recorded in the downstream loop detector (Lint and Zijpp 2003). However, this model performs poorly when the bus speed changes frequently or the speed of change between dual loop detectors cannot well simulate the bus speed in that section. As a result, considering the real-time bus GPS data, this paper proposed a modified model based on Van' model to predict the bus arrival time. As shown in Figure 24, the proposed algorithm is divided into three main parts: (1) Cycle synchronization, (2) Check the exit-point and (3) Travel distance calculation.

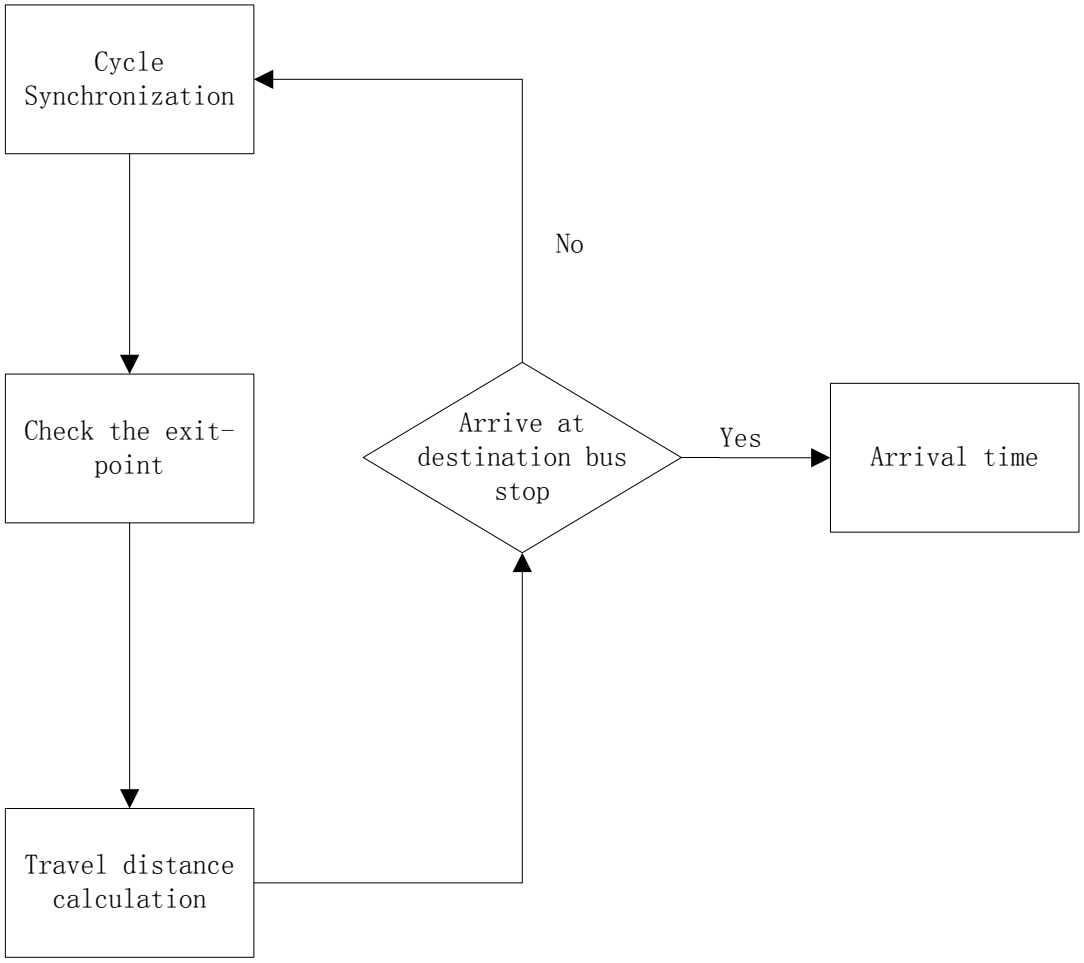

Figure 24 The Workflow of Real-Time Bus Trajectory Reconstruction

Firstly, as the record update cycle for GPS data and loop detector data are not same, the timestamp synchronize process is needed. p is defined as the time period that both the GPS data and loop detector are in the same cycle, if either one turns to be the next cycle, the time period will be defined as p+1. Following the piece-wise linear speed-based (PLSB) model, the speed $v_k(t)$ of a bus k in one section is given by:

$$v_k(\mathrm{t}) = \mathrm{V}(\mathrm{g},\mathrm{p}) + \frac{x_k(t) - x_g}{x_d - x_g} \times (\mathrm{V}(\mathrm{d},\mathrm{p}) - \mathrm{V}(\mathrm{g},\mathrm{p})) \quad (8)$$

where $V(\mathrm{g},\mathrm{p})$ denotes the bus speed from the real time GPS point g during time period p, $\mathrm{V}(\mathrm{d},\mathrm{p})$ denotes the speed recorded in the downstream loop detector d during time period p, x_g denotes

the location of the real time GPS point g, x_d denotes the location of the downstream loop detector d. $x_k(t)$ is the trajectory of bus k as function of time, and it can be given by:

$$x_k(\mathrm{t}) = x_{kip}^E + \left(\frac{\mathrm{V(g,p)}\times(x_d - x_g)}{\mathrm{V(d,p)} - \mathrm{V(g,p)}} + x_{kip}^E - x_u\right) \times \left(e^{\frac{\mathrm{V(d,p)} - \mathrm{V(g,p)}}{x_d - x_g}\times(t - t_{kip}^E)} - 1\right) \quad (9)$$

where x_{kip}^E denotes the entry location of a bus k in section i during time period p, t_{kip}^E denotes the entry time of a bus k in section i during time period p, x_u denotes the location of the upstream loop detector u.

Based on the Eq. (9), the whole path-level trajectory can be reconstructed. As the path is composed of a sequence of sections, the process of checking whether the bus k has exit the section k in time period p is needed. If the condition below matched:

$$x_{kip}^E + \left(\frac{\mathrm{V(g,p)}\times(x_d - x_g)}{\mathrm{V(d,p)} - \mathrm{V(g,p)}} + x_{kip}^E - x_u\right) \times \left(e^{\frac{\mathrm{V(d,p)} - \mathrm{V(g,p)}}{x_d - x_g}\times\left(p - t_{kip}^E\right)} - 1\right) > x_d \quad (10)$$

the exit location x_{kip}^Q and time t_{kip}^Q of bus k in section i during time period p is given by:

$$\begin{cases} x_{kip}^Q = x_d \\ t_{kip}^Q = t_{kip}^E + \frac{x_d - x_g}{\mathrm{V(d,p)} - \mathrm{V(g,p)}} \times \ln\left(\frac{\frac{(x_d - x_g)\times\mathrm{V(g,p)}}{\mathrm{V(d,p)} - \mathrm{V(g,p)}} + x_d - x_u}{\frac{(x_d - x_g)\times\mathrm{V(g,p)}}{\mathrm{V(d,p)} - \mathrm{V(g,p)}} + x_{kip}^E - x_u}\right) \end{cases} \quad (11)$$

Otherwise, the exit location and time of bus k in section i during time period p is given by:

$$\begin{cases} x_{kip}^Q = x_{kip}^E + \left(\frac{(x_d - x_g)\times\mathrm{V(g,p)}}{\mathrm{V(d,p)} - \mathrm{V(g,p)}} + x_{kip}^E - x_u\right) \times \left(e^{\frac{\mathrm{V(d,p)} - \mathrm{V(g,p)}}{x_d - x_g}\times(p - t_{kip}^E)} - 1\right) \\ t_{kip}^Q = p \end{cases} \quad (12)$$

5.2.3 Regression Model

The independent variables selected to develop path-based travel time estimation models were the findings from Chapter 4. Given the above, the general model used to estimate bus travel (and therefore arrival) time from time point i to all downstream time points j is formulated as:

$$T_i = b_0 + b_1 d_{i,j} + b_2 M_p + b_3 A_p + b_4 W_d + b_5 S_n + b_6 V_i + b_7 R_L + b_8 R_C + \varepsilon_i \quad (13)$$

Where T_i is the estimated travel time from time point i to all downstream time points, $d_{i,j}$ is the distance between time point i and time point j, M_p is a binary variable that indicates morning peak, A_p is a binary variable that indicates afternoon peak, W_d is a binary variable that indicates weekend, S_n is a binary variable that indicates moderate snow, V_i is the visibility, R_L is the route length, R_C is the route count, b_0 is the intercept of the travel time estimation model, b_k are the parameters for variables, from 1 to 8, i is the index of origin time points, j is the index of destination time points. ε_i is a random error. It is assumed that $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2, Cov(\varepsilon_i, \varepsilon_j) = 0 \; \forall i \neq j$.

Given the origin time point and time period, the proposed model can estimate the required time to travel the path to every downstream time point and thereby the vehicle arrival time at that time point. All time periods are assigned a value of 1 if present (if the trip started in that time period), and 0 otherwise. Regressions were run both with and without intercepts. All variable notations and their associated coefficients are the same for both types of regression models. The only difference is that models having no intercepts would have their b_0 values equal to zero.

5.3 Results and Discussions

5.3.1 Studied Route

The Route 33 is chosen as the studied route. It is one of the smart bus routes in ETS, and runs between the West Edmonton Mall Transit Center and the Meadows Transit Center. Figure 25 illustrates the map of the Route 33.

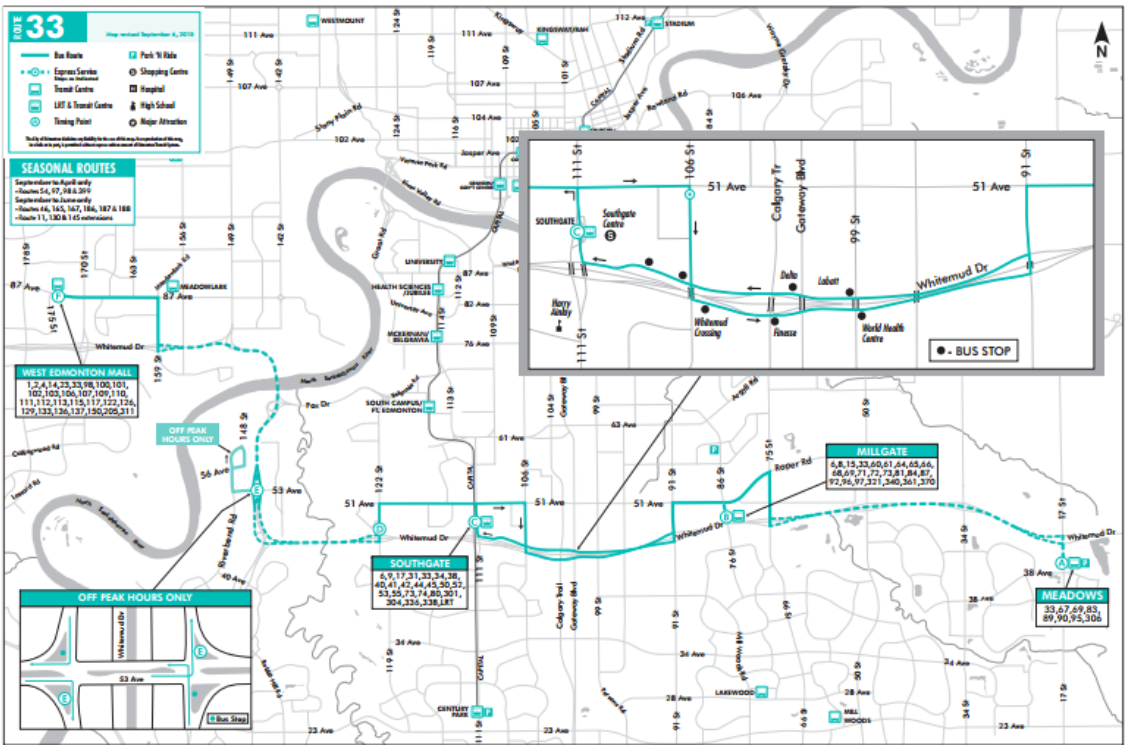

Figure 25 The Map of the Route 33 (City of Edmonton 2015f)

The total length of this route is about 44.5 KM, 25 KM from the Meadows Transit Center to the West Edmonton Mall Transit Center and 19.5 KM from the West Edmonton Mall Transit Center to the Meadows Transit Center. There are 85 stops along the whole route. 37 stops among them are from the Meadows Transit Center to the West Edmonton Mall and 48 stops among them are from the West Edmonton Mall Transit Center to the Meadows Transit Center.

The route section between the bus stop '122 Street & 48 Avenue' to the bus stop '159 Street & Whitemud Drive' is used for the study segment. The total length for this route section is around 7220 meters and the schedule travel time is around 12 minutes. As shown in Figure 26, there are three bus stops along this route section, and the bus arrival time prediction focuses on the time arriving the bus stop '159 Street & Whitemud Drive' from the bus stop '122 Street & 48 Avenue'. As shown in Figure 27, there are 10 loop detectors installed along this route section,

their loop detector IDs are 1018, 1026, 1038, 1040, 1030, 1032, 1035, 1036, 1016, and 1045, respectively.

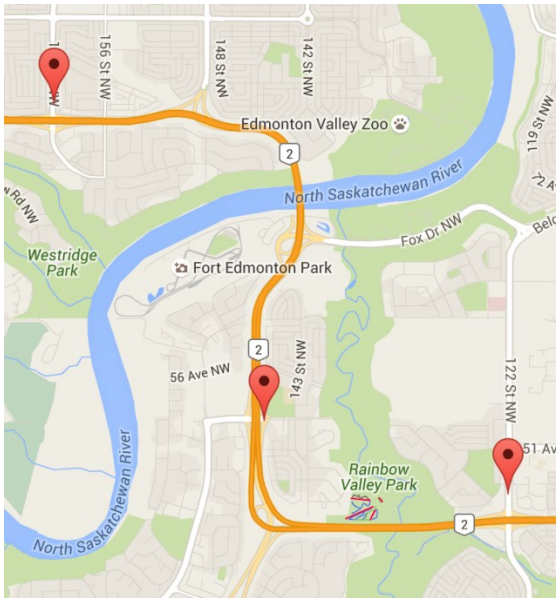

Figure 26 The Study Segment of the Route 33

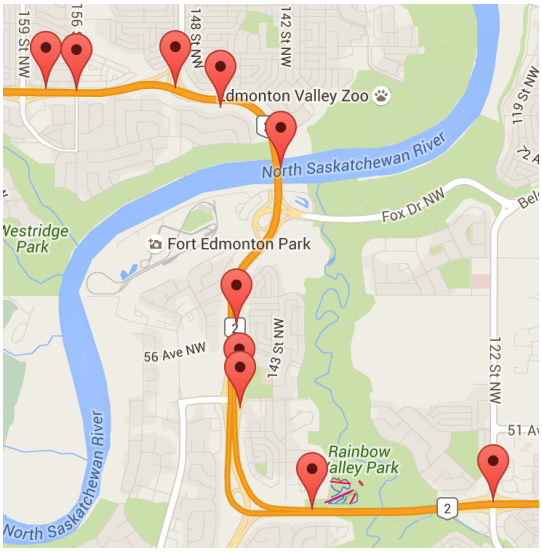

Figure 27 Loop Detector Locations along the Study Segment

5.3.2 Travel Time Analysis

Dec 2nd to Dec 4th is chosen as the study period. For each day, only 9 trips departure from the bus stop '122 Street & 48 Avenue' at 6:23 AM, 6:53 AM, 7:23 AM, 7:53 AM, 8:23 AM,

4:24 PM, 4:54 PM, 5:24 PM, 5:54 PM are selected because the bus will go to the Riverbend community during the other time period, and there are no loop detectors installed in that area. The loop detector data is recorded every 20 seconds for each lane, including the volume, speed, occupancy, loop detector ID and timestamp. The smart bus data can be divided into two types, one is the real time vehicle position data, and it records each bus's real time location data per 30 seconds as below:

<trip>

<tripId>10086861</tripId>

<latitude>53.570408</latitude>

<longitude>113.468445</longitude>

<timeStamp>1448673683</timeStamp>

<speed>60</speed>

<vehicleLabel>4386</vehicleLabel>

</trip>

The other one is the trip update data, which records the bus's departure time and arrival time for each bus stop as below:

<trip>

<tripId>10100278</tripId>

<startDate>20151127</startDate>

<startTime>18:14:00</startTime>

<routeId>33</routeId>

<stopTimeUpdate>

<stopSequence>1</stopSequence>

<arrivalTime>1448673360</arrivalTime>

<depatureTime>1448673368</depatureTime>

<stopId>5001<stopId>

</stopTimeUpdate>

…

<stopTimeUpdate>

<stopSequence>48</stopSequence>

<arrivalTime>1448676000</arrivalTime>

<depatureTime>1448676010</depatureTime>

<stopId>3713<stopId>

</stopTimeUpdate>

<vehicleId>2252</vehicleId>

<vehicleLabel>4620</vehicleLabel>

</trip>

However, due to some sensor errors or data transmission errors, the real time GPS data or trip update data can be missed for some trips, such as the data of 6:53 AM, 7:53 AM, 8:23 AM and 4:24 PM on Dec 2nd, the data of 4:24 PM on Dec 3rd and the data of 4:54 PM on Dec 4th are missed.

Figure 28 shows that the average travel time per trip in these three days are similar and they are 526, 599, and 608 seconds, respectively. The travel time ranges for all trips in these three days are different, especially between Dec 2nd, Dec 3rd and Dec 4th.

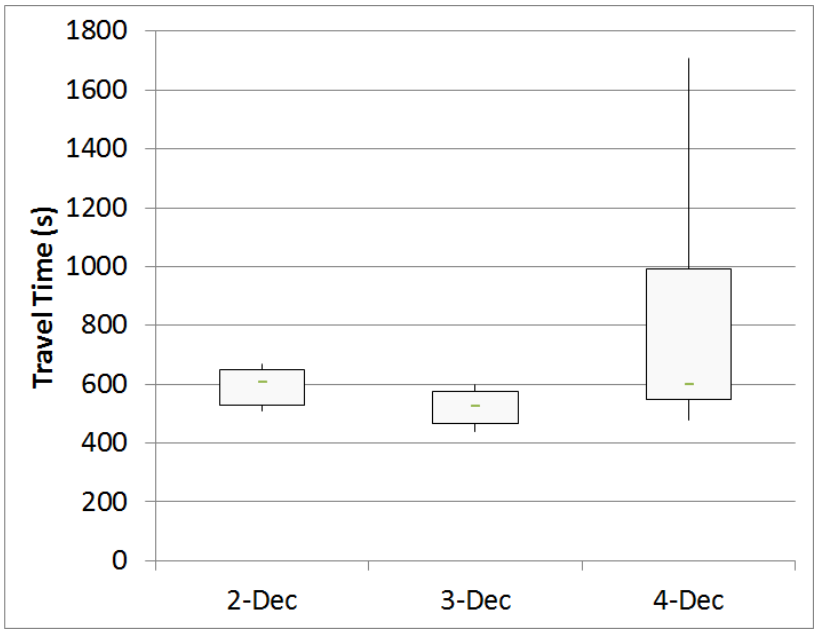

Figure 28 Box Plot of Travel Time from Dec 2nd to Dec 4th

As shown in Figure 29, the reason that the travel time ranges of Dec 4th is bigger than the other days is that Dec 4th is Friday, and there is usually a heavy traffic congestion during the PM peak hours of Friday. While for the AM periods in these three days, the average travel time is similar, they are 638, 535, and 581 seconds, respectively and the average time for PM periods from Dec 2nd to Dec 4th are 572, 519 and 1545 seconds, respectively.

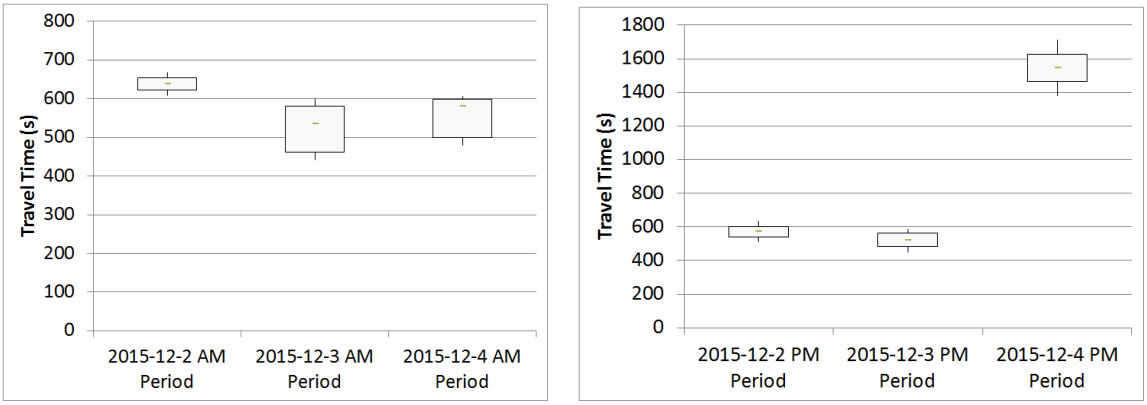

Figure 29 Box Plot of AM/PM Travel Time from Dec 2nd to Dec 4th

5.3.3 Bus Arrival Time Prediction

5.3.3.1 Results based on GPS Data

Using the bus arrival time prediction model, the bus trajectories can be reconstructed, then the arrival time from the bus stop '122 Street & 48 Avenue' to the bus stop '159 Street & Whitemud Drive' can be figured out through the reconstructed trajectories. As shown in Figure 30, the predicted and the GPS trajectories of all trips on Dec 2nd are plotted, and it shows a satisfactory agreement between the predicted and the GPS trajectories. In the figure, there is a horizontal line around 3000 meters for each trajectory, which means the bus stop 'Whitemud Drive NB & 53 Avenue', and because this bust stop is a time point bus stop, each bus needs to wait until the schedule departure time. The schedule travel time from the departure time of bus stop '122 Street & 48 Avenue' to the departure time of bus stop 'Whitemud Drive NB & 53 Avenue' is 6 minutes. However, as shown in Figure 30, many buses don't follow this rule, and this is related to the driver behavior. Also, buses will accelerate from the departure bus stop and decelerate to the arrival bus stop, in this study, the acceleration rate is calculated from the historical data from Dec 2nd to Dec 4th, and the estimated acceleration rate is 2.5 m/s^2. For the deceleration process, it is assumed that each bus will start to decelerate 50 meters to the bus stop in this study. The result of the reference arrival time and the predicted arrival time at bus stop '159 Street & Whitemud Drive' is showed in Table 3.

Table 2 Arrival Time Result on Dec 2nd

Departure time	Reference travel time (s)	Predicted travel time (s)
6:22:02	608	593
7:22:31	669	677
16:53:01	633	644
17:26:05	511	533
17:54:35	544	559

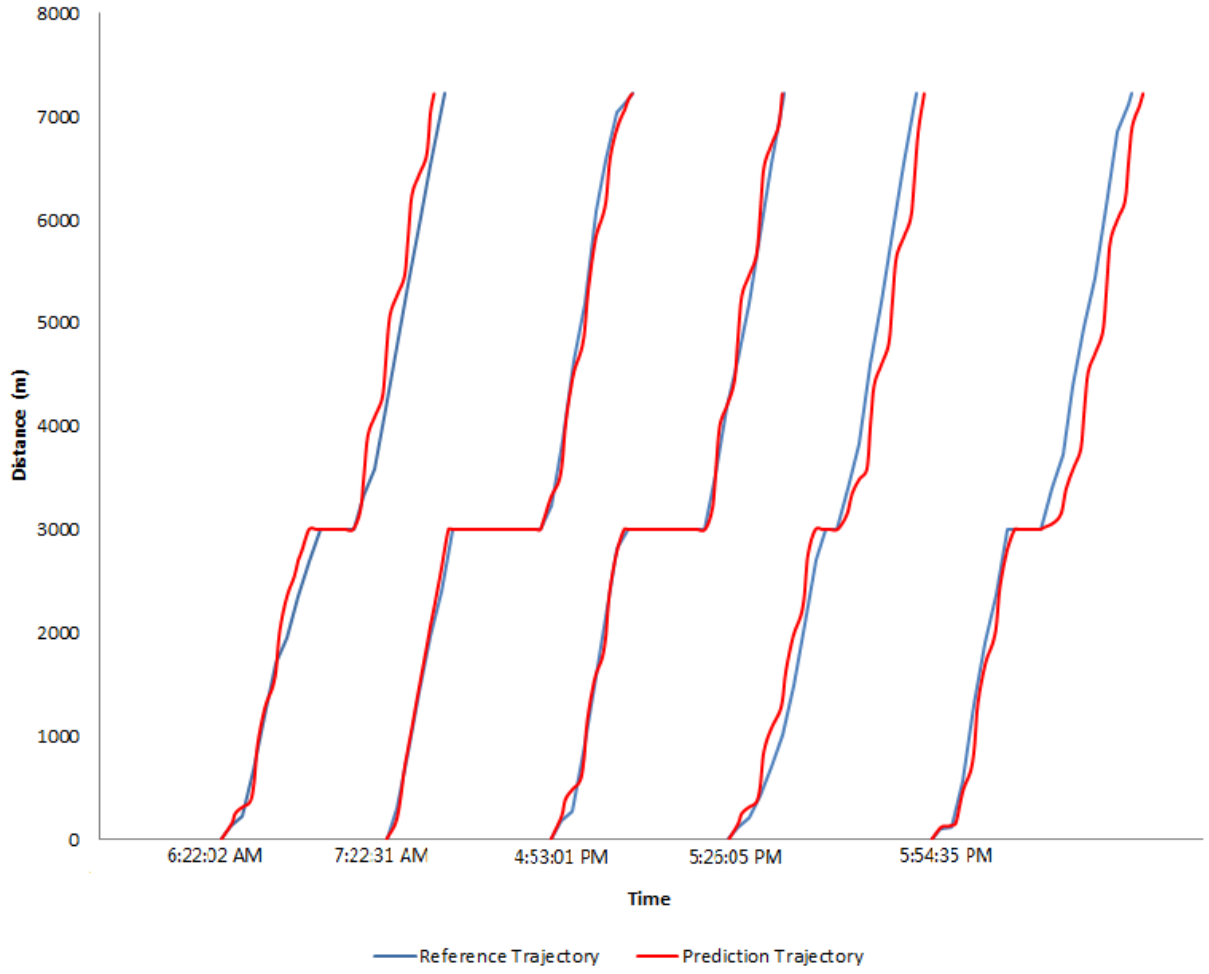

Figure 30 Prediction Trajectories vs Reference Trajectories on Dec 2nd

Figure 31 shows the predicted and the reference trajectories of all trips on Dec 3rd, it also presents a satisfactory agreement between the predicted and the reference trajectories. The result of the reference arrival time and the predicted arrival time at bus stop '159 Street & Whitemud Drive' is showed in Table 4.

Table 3 Arrival Time Result on Dec 3rd

Departure time	Reference travel time (s)	Predicted travel time (s)
6:22:36	600	613
7:54:54	484	481
7:25:40	535	552
7:57:55	441	459
8:25:32	563	574
16:55:03	450	460
17:26:31	517	539

17:53:44	588	584

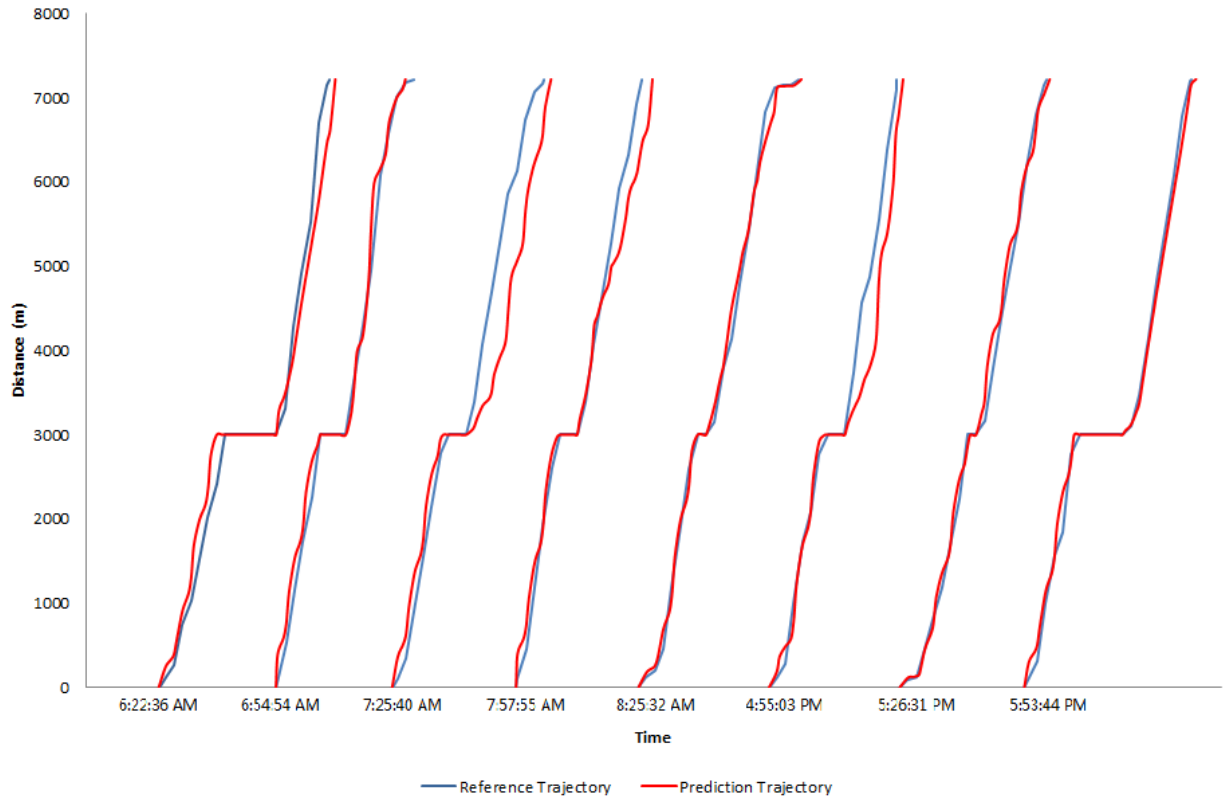

Figure 31 Prediction Trajectories vs Reference Trajectories on Dec 3rd

Figure 32 shows the predicted and the reference trajectories of all trips on Dec 4th. As mentioned above, there is a heavy congestion in the PM periods on Dec 4th, and the trajectories in AM periods are steeper than in PM periods. The result of the reference arrival time and the predicted arrival time at bus stop '159 Street & Whitemud Drive' is showed in Table 5.

Table 4 Arrival Time Result on Dec 4th

Departure time	Reference travel time (s)	Predicted travel time (s)
6:29:27	581	592
6:52:27	593	613
7:28:01	606	590
7:52:49	519	538
8:53:57	479	500
16:32:07	1381	1369
17:29:01	1928	1912
17:56:02	1710	1736

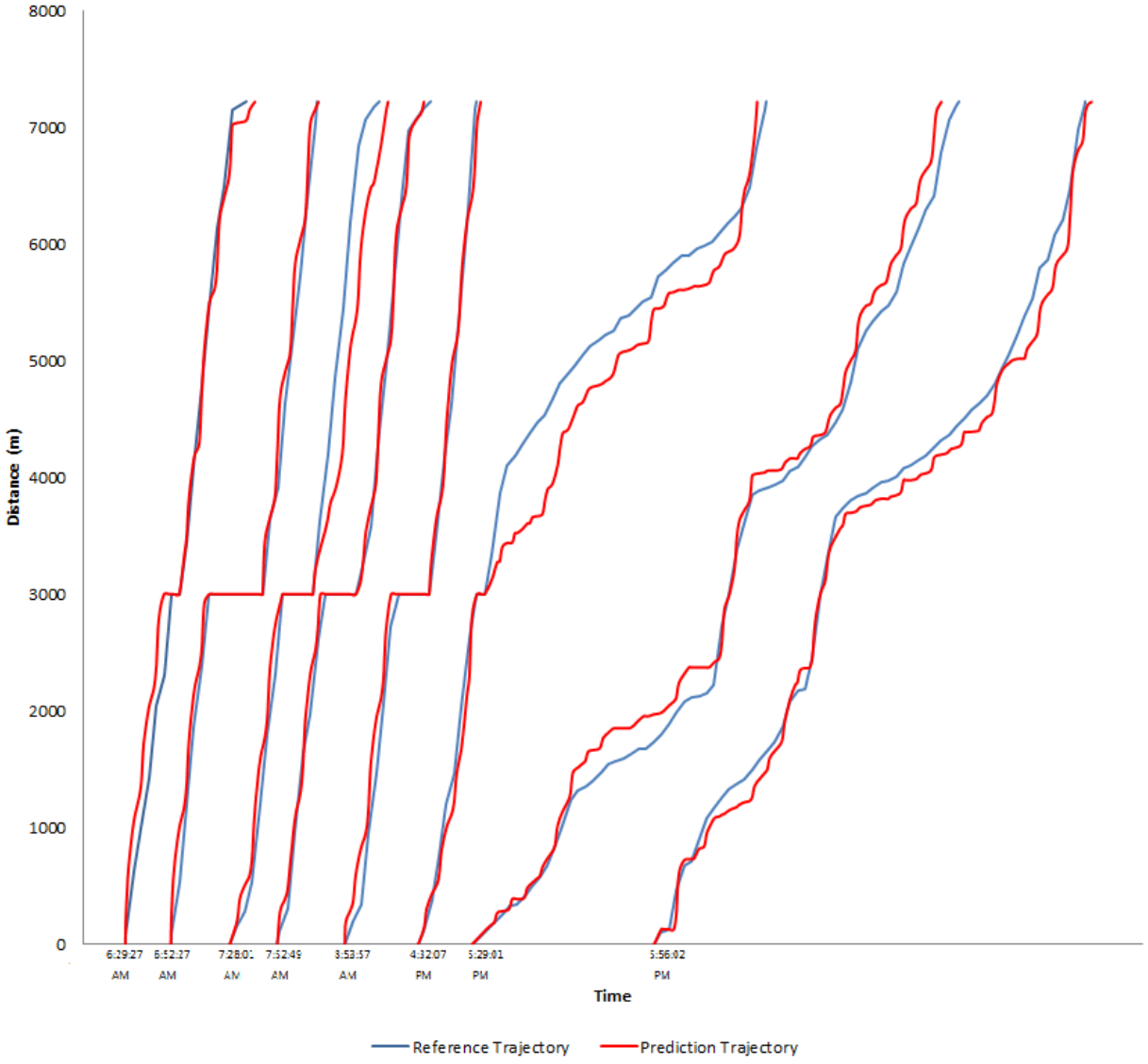

Figure 32 Prediction Trajectories vs Reference Trajectories on Dec 4th

To validate the proposed algorithm and to evaluate the agreements between the predicted bus arrival time and the actual bus arrival time quantitatively, two evaluation measures are used: (1) mean absolute percentage error (MAPE) and root mean squared Error (RMSE). MAPE and RMSE are calculated by the following two equations:

$$MAPE = \frac{1}{k}\sum_{i}^{k}\frac{|t_i - t_a|}{t_a} \times 100\% \quad (14)$$

$$RMSE = \sqrt{\frac{1}{k}\sum_{i}^{k}(t_i - t_a)^2} \quad (15)$$

where t_i denotes the predicted bus arrival time, t_a denotes the actual bus arrival time and k is the number of predictions.

Table 6 shows the goodness of fit of the proposed algorithm. The table shows that the proposed algorithm could successfully estimate the vehicle trajectories and predict the bus arrival time.

Table 5 Goodness of Fit Statistics

	MAPE (%)	**RMSE (s)**
Three days	2.41	16
Dec 2nd	2.45	15
Dec 3rd	2.39	14
Dec 4th	2.40	18
Under congestion	1.07	19
Not under congestion	2.63	15

5.3.3.2 Results based on Regression Model

The plot of actual versus estimated bus travel time is presented in Figure 33. It substantiates visually the linear relationship of the dependent variable with all independent variables that are used in the models. The overall model statistics are shown in Table 7.

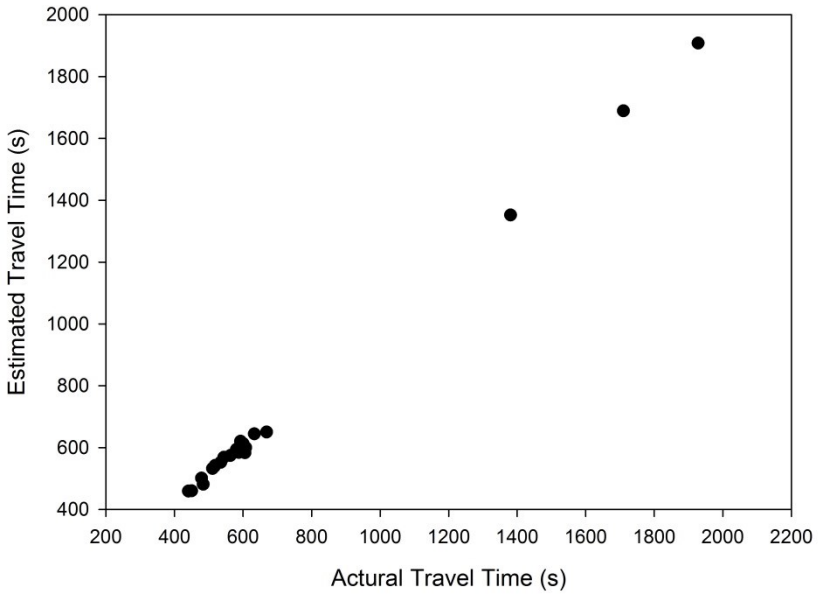

Figure 33 Estimated Versus Actual Travel Time

Table 6 Statistics of Bus Travel Time Estimation Models

Parameters	b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8
Value	1.11	2.61	-0.87	3.56	-1.05	2.57	0.57	1.23	1.67
R-Sq: 0.96; RMSE: 18s; p value: <0.0001									

5.3.3.3 Results Comparison

The MAPE was used as the measure of effectiveness in this paper. Table 8 shows the average MAPE for the two models. It is hypothesized that the congestion reduces the variability in travel times and this makes the models more accurate for this time period. The use of real-time schedule adherence data for the GPS data-based model did significantly improve the results. For this test bed the GPS data-based model gave superior results, in terms of MAPE, in comparison with the multi-linear regression results.

Table 7 Average MAPE of the Prediction Models

		MAPE (%)
Under congestion	GPS data-based model	1.07
	Regression model	1.09
Not under congestion	GPS data-based model	2.63
	Regression model	3.25

CHAPTER 6 CONLUSIONS

6.1 Research Summary

The improvement of the efficiency of Transit Service Reliability (TSR) is critical to attract additional ridership and increase the satisfaction of public transit, which is the key to developing the future environmentally conscious and sustainable transportation system. The availability of amounts of multi-source data has gradually shifted the public transit system into a data-rich environment, and makes it realizable to do further TSR analysis and predict bus arrival time more preciously.

Five kinds of data were used in this study: VDS data, APC data, General Transit Feed Specification (GTFS) data, weather data and road network Geographic Information System (GIS) data. As those multi-source data were maintained by different branches with diversity of formats, a data integration process was implemented in terms of three parts: (a) spatial data representation, (b) Uniform Location Reference System (LRS) establishment, (c) topology relationship establishment. Thereafter, each loop detector, bus stop, smart bus GPS point, bus route, intersection and road segment had been abstracted into their corresponding GIS object with a location coordinate in a uniform LRS, and the topology among them was set as well.

An understanding of bus service reliability was necessary to develop strategies that help transit agencies provide better services. This study identified and quantified the temporal characteristics, weather characteristics and operating characteristics affecting bus arrival time variability. For the temporal characteristics, the result showed that the likelihood of both early and late arrivals increased in the afternoon period, whereas in the morning period the likelihood of early arrival increased and the likelihood of late arrival decreased. In the weekend, the

likelihood of early arrival increased and the likelihood of late arrival decreased. For the weather characteristics, the result showed that the likelihood of late arrivals increased significantly under the moderate snow condition. Moreover, late arrivals become more likely with lower visibility. For the operating characteristics, the result showed that a long route resulted in increases in the likelihood of both early and late arrivals. The probability of late arrivals tended to increase as buses progress toward a route's terminal point.

Bus arrival time information was useful for passengers to reduce the waiting time at bus stops or to make reasonable travel arrangements before making a trip. This study proposed a bus trajectory reconstructive model to predict the bus arrival time using GPS data and VDS data. After projecting the stored GPS records on the transit network and the road network, the bus trajectory was reconstructed following three steps: cycle synchronization, checking the exit-point, and travel distance calculation. In addition, regression model based on the factors found in the first part of this study was applied as a comparison. The results indicated that the proposed algorithm was capable of achieving satisfactory accuracy in predicting bus arrival time. The use of real-time schedule adherence data for the GPS data-based model did significantly improve the results. For this test bed the GPS data-based model gave superior results, in terms of MAPE, in comparison with the multi-linear regression results.

6.2 Future Work

While the results are encouraging, there are still a number of endeavors that should be made to enhance both the depth and width of the proposed work.

First, it is hypothesized that if real-time accident and incident information is available, arrival time predictions will be improved. As the applications of connected vehicle technology

and social media are developing quickly, this will not be an impediment in the near future. In addition, Due to the incomplete of the APC data, only some parts of the APC data are selected in the evaluation of factors affecting bus on-time performance. In the future, when the APC data is complete and without error, more factors from the APC data can be added to do further analysis, like the onboarding and alighting numbers, etc.

More experiments are needed to evaluate the model performance thoroughly. This study only predicted the bus arrival time at time-point bus stops and assumed the drivers always following the time-point schedule. However, randomness and driver behavior factor should be taken into account to estimate the bus dwell time. Therefore, the model can be extended to predict the bus arrival time along a whole bus route.

Future research could be developed in how bus schedules can be further optimized to be robust knowing that the snow weather effects exist. For example, a dynamic timetable has substantial potential to improve TSR during adverse weather conditions.

REFERENCES

Abkowitz, Mark, and Israel Engelstein. 1984. "Methods for Maintaining Transit Service Regularity." *Transportation Research Record: Journal of the Transportation Research Board* 361: 1–8.

Albright, E., and M. Figliozzi. 2012. "Factors Influencing Effectiveness of Transit Signal Priority and Late-Bus Recovery at Signalized-Intersection Level." *Transportation Research Record: Journal of the Transportation Research Board* 2311: 186–97.

Arhin, Stephen, and P.E. PTOE. 2013. "Evaluation of Bus Transit Reliability in the District of Columbia."

Bae, S., and P. Kachroo. 1995. "Proactive Travel Time Predictions under Interrupted Flow Conditions." In *Vehicle Navigation and Information Systems Conference*, 179–86.

Bates, J. W. 1986. "Definition O F Practices for Bus Transit On-Time Performance: Preliminary Study." *Transportation Research Board, National Research Council* 300: 1–5.

Beijing Transportation Research Center. 2010. "Beijing Transportation Smart Card Usage Survey."

Bertini, R.L., and A.M. El-Geneidy. 2004. "Modeling Transit Trip Time Using Archived Bus Dispatch System Data." *Journal of Transportation Engineering* 130 (1): 56–67.

Boilé, Maria P. 2001. "Estimating Technical and Scale Inefficiencies of Public Transit Systems." *Journal of Transportation Engineering* 127 (3): 187–94.

Cham, Laura Celia. 2006. "Understanding Bus Service Reliability: A Practical Framework Using AVL/APC Data."

Chen, Xumei, Lei Yu, Yushi Zhang, and Jifu Guo. 2009. "Analyzing Urban Bus Service Reliability at the Stop, Route, and Network Levels." *Transportation Research Part A: Policy and Practice* 43 (8): 722–34.

Chien, S. I. J., Y. Ding, and C. Wei. 2002. "Dynamic Bus Arrival Time Prediction with Artificial Neural Networks." *Journal of Transportation Engineering* 128 (4): 29–438.

City of Edmonton. 2015a. "ETS Bus Schedule GTFS Data Feed - Routes." https://data.edmonton.ca/Transit/ETS-Bus-Schedule-GTFS-Data-Feed-Routes/d577-xky7.

City of Edmonton. 2015b. "ETS Bus Schedule GTFS Data Feed - Stops." https://data.edmonton.ca/Transit/ETS-Bus-Schedule-GTFS-Data-Feed-Stops/4vt2-8zrq.

City of Edmonton. 2015c. "ETS Bus Schedule GTFS Data Feed - Trips." https://data.edmonton.ca/Transit/ETS-Bus-Schedule-GTFS-Data-Feed-Trips/ctwr-tvrd.

City of Edmonton. 2015d. "ETS Day Map." http://www.edmonton.ca/transportation/transit/ETS_Day_Map_Fall_2015.pdf.

City of Edmonton. 2015e. "Real Time Vehicle Position GTFS (PB File)." https://data.edmonton.ca/Transit/Real-Time-Vehicle-Position-GTFS-PB-File-/7qed-k2fc.

City of Edmonton. 2015f. "Route 33 Schedule and Map." http://webdocs.edmonton.ca/transit/route_schedules_and_maps/current/RT033.pdf.

City of Edmonton. 2015g. "Smart Bus Project." http://www.edmonton.ca/ets/transit_projects/smart-bus-initiative.aspx.

Consulting, ICF. 2003. "Strategies for Increasing the Effectiveness of Commuter Benefits Programs."

Crout, D. 2007. “Accuracy and Precision of the Transit Tracker System.” *Transportation Research Record: Journal of the Transportation Research Board* 1992: 93–100.

Diab, E., and A. El-Geneidy. 2013. “Variation in Bus Transit Service: Understanding the Impacts of Various Improvement Strategies on Transit Service Reliability.” *Public Transport* 4 (3): 209–31.

Diab, E.I., and A.M. El-Geneidy. 2012. “Understanding the Impacts of a Combination of Service Improvement Strategies on Bus Running Time and Passenger’s Perception.” *Transportation Research Part A: Policy and Practice* 46 (3): 614–25.

Dueker, K.J., T.J. Kimpel, J.G. Strathman, and S. Callas. 2004. “Determinants of Bus Dwell Time.” *Journal of Public Transportation* 7 (1): 21–40.

El-Geneidy, A.M., J. Horning, and K.J. Krizek. 2011. “Analyzing Transit Service Reliability Using Detailed Data From Automatic Vehicular Locator Systems.” *Journal of Advanced Transportation* 45 (1): 66–79.

El-Geneidy, A.M., and J. Surprenant-Legault. 2010. “Limited-Stop Bus Service: An Evaluation of An Implementation Strategy.” *Public Transport* 2 (4): 291–306.

El-Geneidy, Ahmed M, John Hourdos, and Jessica Horning. 2009. “Bus Transit Service Planning and Operations in a Competitive Environment.” *Transportation Research Record: Journal of the Transportation Research Board* 12: 39–60.

Environment Canada. 2015. “Canadian Climate Data.” http://climate.weather.gc.ca/.

Figliozzi, M., and W. Feng. 2012. “A Study of Headway Maintenance for Bus Routes: Causes and Effects of ‘Bus Bunching’ in Extensive and Congested Service Areas.” Vol. OTREC-

RR-1.

Frechette, L. A., and A. M. Khan. 1998. “Bayesian Regression-Based Urban Traffic Models.” *Transportation Research Record: Journal of the Transportation Research Board* 1644: 157–65.

Furth, P., B. Hemily, T. H. J. Muller, and J. Strathman. 2006. “Using Archived AVL-APC Data to Improve Transit Performance and Management.”

Google Inc. 2015a. “Realtime Transit.” https://developers.google.com/transit/gtfs-realtime/?hl=en.

Google Inc. 2015b. “Static Transit.” https://developers.google.com/transit/gtfs/?hl=en.

Guerts, Sabine A., Wilmar B. Schaufeli, and Bram P. Buunk. 1993. “Social Comparison, Inequity, and Absenteeism among Bus Drivers.” *European Work and Organizational Psychologist* 3 (3): 191–203.

Gurmu, Zegeye K, and Wei Fan. 2014. “Artificial Neural Network Travel Time Prediction Model for Buses Using Only GPS Data.” *Journal of Public Transportation* 17 (2): 45–65.

Hensher, D.A., P. Stopher, and P. Bullock. 2003. “Service Quality - Developing a Service Quality Index in the Provision of Commercial Bus Contracts.” *Transportation Research Part A: Policy and Practice* 37 (6): 499–517.

J., An, Y. Liu, and X. Yang. 2014. “Measuring Route-Level Passenger Perceived Transit Service Reliability with an Agent-Based Simulation Approach.” *Transportation Research Record: Journal of the Transportation Research Board* 2415: 48–58.

Jeong, R., and L. R. Rilett. 2004. “Bus Arrival Time Prediction Using Artificial Neural Network

Model." In *IEEE Intelligent Transportation Systems Conference*, 988–93.

Kittelson & Associates, Inc. 2013. "Transit Capacity and Quality of Service Manual."

Kjmpel, Thomas Jeffrey. 2001. "Time Point-Level Analysis of Passenger Demand and Transit Service Reliability."

Levinson, H S. 1991. "Supervision Strategies for Improved Reliability of Bus Routes." In *Transportation Research Board*, 74.

Levinson, Herbert S. 2005. "The Reliability of Transit Service: An Historical Perspective." *Journal of Urban Technology* 12 (1): 99–118.

Lint, J.W.C. Van, and N.J. Van der Zijpp. 2003. "Improving A Travel Time Estimation Algorithm by Using Dual Loop Detectors." *Transportation Research Record* 1855: 41–48.

McKenzie, B., and M. Rapino. 2009. "Commuting in United States: 2009, American Community Survey Reports."

McKnight, C., H. Levinson, K. Ozbay, C. Kamga, and R. Paaswell. 2004. "Impact of Traffic Congestion on Bus Travel Time in Northern New Jersey." *Transportation Research Record: Journal of the Transportation Research Board* 1884: 27–35.

Mesbah, M., G. Currie, and N. Peñafiel Prohens. 2014. "Effect of Daylight on Reliability of Transit Service: Case Study of Melbourne Tram Network, Australia." In *93rd Transport Research Board Annual Meeting (TRB 2014)*.

Mesbah, M., J. Lin, and G. Currie. 2015. "'Weather' Transit Is Reliable? Using AVL Data to Explore Tram Performance in Melbourne, Australia." *Journal of Traffic and Transportation Engineering (English Edition)* 2 (3): 125–35.

Murray, A.T., and X. Wu. 2003. “Accessibility Tradeoffs in Public Transit Planning.” *Journal of Geographical Systems* 5 (1): 93–107.

Patnaik, J., S. Chien, and A. Bladihas. 2004. “Estimation of Bus Arrival Times Using APC Data.” *Journal of Public Transportation* 7 (1): 1–20.

Shalaby, A., and A. Farhan. 2003. “Bus Travel Time Prediction for Dynamic Operations Control and Passenger Information Systems.” In *82nd Annual Meeting of the Transportation Research Board*, 1–16.

Slavin, C., W. Feng, M. Figliozzi, and P. Koonce. 2013. “A Statistical Study of the Impacts of SCATS Adaptive Traffic Signal Control on Traffic and Transit Performance.” *Transportation Research Record: Journal of the Transportation Research Board* 2356: 117–26.

Sterman, Brian P., and Joseph L. Schofer. 1976. “Factors Affecting Reliability of Urban Bus Services.” *Transportation Engineering Journal of ASCE* 102 (1): 147–59.

Strathman, J.G., T.J. Kimpel, K.J. Dueker, R.L. Gerhart, and S. Callas. 2002. “Evaluation of Transit Operations: Data Applications of Tri-Met’s Automated Bus Dispatching System.” *Transportation* 29 (3): 321–45.

Tétreault, P.R., and A.M. El-Geneidy. 2010. “Estimating Bus Run Times for New Limited-Stop Service Using Archived AVL and APC Data.” *Transportation Research Part A: Policy and Practice* 44 (6): 390–402.

Texas Transportation Institute. 2005. “2005 Urban Mobility Report.”

Van Oort, Niels, and Robert van Nes. 2009. “Line Length Versus Operational Reliability

Network Design Dilemma in Urban Public Transportation." *Transportation Research Record: Journal of the Transportation Research Board* 2112: 104–10.

Vanajakshi, L., S. C. Subramanian, and R. Sivanandan. 2009. "Travel Time Prediction under Heterogeneous Traffic Conditions Using Global Positioning System Data from Buses." *Intelligent Transport Systems* 3 (1): 1–9.

Wall, Z., and D. J. Dailey. 1999. "An Algorithm for Predicting the Arrival Time of Mass Transit Vehicles Using Automatic Vehicle Location Data." In *78th Annual Meeting of the Transportation Research Board*, 1–11.

Xuan, Yiguang, Juan Argote, and Carlos F. Daganzo. 2011. "Dynamic Bus Holding Strategies for Schedule Reliability: Optimal Linear Control and Performance Analysis." *Transportation Research Part B: Methodological* 45 (10): 1831–45.