

**Bus Bunching Prediction and Transit Route Demand
Estimation Using Automatic Vehicle Location Data**

SUN WENZHE

Abstract

The availability of massive passive transit data has provided a variety of feasibilities for the public transit operators to deliver better service that is of high reliability and characterizes users' demand. Among these data sources, Automatic Vehicle Location (AVL), Automatic Fare Collection (AFC) and Automatic Passenger Counting (APC) data are commonly used, and they are playing significant roles in different aspects of improving transit services. AVL data contain the spatiotemporal coordinates of the transit vehicles which not only offer service reliability metrics, but also indicate the general traffic condition. AFC and APC data are presenting passenger demand patterns, including boarding flows, alighting flows, passenger loads and, in case of AFC only, origin-destination (OD) flows.

AVL data are easier to obtain compared to AFC data. Existing literature shows the effectiveness of AVL data in predicting future arrival times and headways, and providing useful inputs for the models to control bus vehicles in terms of operation velocity and headway in real time. However, the methodology for using bus AVL data to estimate demand has been fairly undeveloped.

This research has two main objectives. Firstly, it aims to illustrate the characteristics of bus AVL data, overview the conventional applications which are using this data source to make predictions on various items of operators' concern and further extend the existing literature by taking bus bunching as the prediction object. Secondly, it proposes an innovative idea and a novel methodology to estimate bus route OD flows using the bus AVL data as the main data source. It expands the potential of bus AVL data in inferring passenger demand beyond the design purpose of AVL system. This can help the operators having limited access to AFC and APC data.

In the first part of the dissertation, a logistic regression model is developed to compute the probability that a bus is going to be caught in bunching instead of answering the question in a deterministic yes-or-no form based on the exact value predicted for headway. The prediction performance of the proposed method is compared with two headway-based methods using linear regression and support vector machine. Bus AVL data from Kyoto City, Japan, are used for the case study. The advantages of the probabilistic method are illustrated by Receiver Operator Characteristic (ROC) analysis in which the ROC curve is interpreted as an efficient front providing the operators with trade-off options.

In the second part of the dissertation, the feasibility of estimating passenger OD flows for a bus transit route using AVL data is explored. Dwell time models are used to build connections between the information extracted from bus AVL data and passenger OD flows. A modified gravity model is applied to reduce the number of unknown parameters. Bayesian inference and Markov Chain Monte Carlo (MCMC) methods are implemented to estimate the mean and confidence intervals for the unknown parameters. The methodology is validated with bus AVL data from Shizuoka City, Japan where AFC data for nearly all passengers in the same observational period are available and thus can be taken as the ground truth. It is found in the case study that the estimation performance derived by using bus AVL data matches that using independent passenger boarding/alighting counts. Furthermore, additional input data that roughly characterize the importance of bus stops improve the performance of OD estimation.

Finally, complex dwell time models considering the effect of various methods of payment, in-vehicle crowding and bus bunching are discussed. It is shown that the estimated demand might be biased in case a too simplistic dwell time model is applied and bus bunching and crowding are frequently occurring.

Keywords: Automatic Vehicle Location data; Bus bunching prediction; Origin-destination matrix; Bayesian inference; Bus dwell time

Acknowledgements

At the end of the long hard journey after pursuing the master's degree and PhD in Japan for more than five years, I am truly grateful to the people who I am luckily connected with for offering me the kind help in any form.

First of all, I would like to thank my supervisor Dr. Jan-Dirk Schmöcker, who has been playing a variety of roles and guiding me in these years, including supervisor, mentor, co-author, boss and academia icon as well. You make me feel that pursuing PhD in this lab is one of the best choices I have made in my life. It did not occur to me that I would go this far with respect to research when I started my study in Kyoto University as a master's course student. And enlightened by the interactive mentorship you provide me, I am now confident that I can go even further. You teach me how to determine the research direction, to find interesting and meaningful topics, to write the research paper, to guide the students, and most importantly the right attitude towards research. With these gifts I can try to build up an academic career, and feel motivated to pass them to my future students.

I also thank Prof. Yamada and Prof. Fujii in Kyoto University, who are the committee member of my PhD dissertation and give me insightful comments in the seminars and discussions. And thanks to the current faculty of Intelligent Transportation Systems Laboratory, Prof. Yamada and Dr. Schmöcker again, Dr. Nakao and lab secretary, Ms. Nishikawa, for giving me a lot of help in research and other aspects.

Sincere thanks to previous lab faculty, Prof. Uno in Kyoto University, Dr. Nakamura currently in Nagoya University and Dr. Yamazaki who currently goes to the industry. I probably even failed to complete the master's course if without your help in the initial stage when I started this journey. Also thanks to the previous lab secretaries, Ms. Shii, Ms. Nishimura and Ms. Nagata for melting my inside uncertainties with your warm hearts, helping me in landing the lives in Japan, administrative procedures and so forth. Although you left our lab for various reasons, I will always keep this in my mind.

Thanks to the past and present students in our lab. The night for the first time I arrived at Kansai Airport, Japan is like yesterday. Due to technical problems, it took me almost two hours to complete

the immigration procedure. I was touched when seeing Fujii-kun and Senda-kun waiting for two hours and holding a paperboard with my name. I still keep the paperboard in my drawer, which I think also the souvenir of all the lovely lab members I meet in this journey, classmates, seniors, juniors, colleagues and friends. Apology for my not naming you one by one here.

Genuine thanks to Lotte foundation for the financial support in recent two years. This scholarship is indeed timely help to me. Also thank you for the networking events that provide me the opportunity to connect with so many excellent international students in Japan, and the lifelong chocolates.

I acknowledge the financial support from the Kyoto Innovative Transportation Research Organization for part of my research in 2018 and 2019. Thanks to Hitachi Laboratory (Hitach R&D Group) in Kyoto University for the financial support and notably Dr. Fukuda for the valuable advices on my research which greatly inspire me.

Finally, I would like to thank my parents, Yunfang Sun and Huiying Xu. I feel speechless regarding the words I should send to you after being absent from your side in so many days and nights. The accomplishment of completing this dissertation undoubtedly also belongs to you.

Preface

Parts of this dissertation have been published in journals, presented in conferences or else submitted for review are as follows

Journal papers

- I. **Sun, W.**, Schmöcker, J.-D., & Nakamura, T. (2020). On the trade-off between sensitivity and specificity in bus bunching prediction. *Journal of Intelligent Transportation Systems* (available online). **(Chapter 3)**

Conference presentations

- I. **Sun, W.**, & Schmöcker, J.-D. (2019). A method to classify bus bunching events using AVL data. Transportation Research Board 98th Annual Meeting, Washington DC, USA. **(Chapter 3)**
- II. **Sun, W.**, & Schmöcker, J.-D. (2019). Using AVL data to predict public transport bunching. 60th Autumn Conference of Committee of Infrastructure Planning and Management, Japan Society of Civil Engineers. Toyama, Japan **(Chapter 3)**
- III. **Sun, W.**, Schmöcker, J.-D., Nakamura, T., & Shimamoto, H. (2018). Bus bunching prediction based on logistic regression considering rare event bias. CASPT 2018 and TransitData 2018, Brisbane, Australia. **(Chapter 3)**
- IV. **Sun, W.**, Schmöcker, J.-D., & Fukuda, K. (2019). Real-time estimation of bus passenger OD patterns based on AVL data. TransitData 2019, Paris, France. **(Chapter 4)**

Under review/ To be submitted

- II. **Sun, W.**, Schmöcker, J.-D., & Fukuda, K. (2020). Estimation of transit route origin-destination flows using bus AVL data. *Transportation research part C: Emerging Technologies* (under review). **(Chapter 4)**
- III. **Sun, W.**, Schmöcker, J.-D., Fukuda, K., & Nakamura, T. (2020). Estimation of passenger flows through AVL data with different dwell time functions. TransitData 2020, Toronto, Canada (under review). **(Chapter 5)**

Table of Contents

Chapter 1	Introduction.....	1
1.1	Background and research motivation	1
1.1.1	Passive transit data.....	1
1.1.2	The characteristics of AVL, APC and AFC data.....	1
1.1.3	Understanding the pros and cons of AVL data.....	2
1.1.4	Bus bunching prediction: leveraging the strength of AVL data	3
1.1.5	Route demand estimation: bridging the gap between AVL data and APC/AFC data	4
1.2	Research objectives	4
1.3	Research contributions.....	6
1.4	Outline of the dissertation.....	7
Chapter 2	Literature review	9
2.1	Overview of bus AVL data.....	9
2.2	Studies using bus AVL data as the main data source	14
2.2.1	In bus network assessment and planning.....	14
2.2.2	In bus operation	17
2.3	Studies combining bus AVL data and APC/AFC data	18
2.4	Bus bunching problem.....	19
2.4.1	Bus control strategies.....	19
2.4.2	Bunching prediction	24
2.5	OD estimation.....	27
2.6	Passenger load estimation and prediction	29
Chapter 3	Bus bunching prediction	31

3.1 Introduction	31
3.2 Bus bunching prediction	32
3.2.1 The identification of bus bunching event	32
3.2.2 Variable selection	33
3.2.3 Logistic regression.....	34
3.2.4 Rare events bias.....	34
3.2.5 Linear regression and SVM as benchmark solutions.....	36
3.3 Data processing.....	36
3.4 Headway prediction results.....	39
3.4.1 Linear regression results.....	39
3.4.2 Support vector machine	40
3.4.3 Performance evaluation index	41
3.4.4 Performance comparison	41
3.5 Bunching prediction.....	43
3.5.1 Logistic regression results	43
3.5.2 Performance evaluation index	46
3.5.3 Performance comparison	47
3.6 Discussion on the trade-off between sensitivity and specificity	52
3.7 Summary.....	54
Chapter 4 Demand estimation using bus AVL data	56
4.1 Introduction	56
4.2 Model framework	57
4.3 Basic dwell time models	62
4.4 Bayesian inference.....	63

4.5 Solution algorithms.....	67
4.6 Data processing.....	73
4.6.1 Dwell time, headway and vehicle activity time	73
4.6.2 Boarding/alighting time per passenger	75
4.7 Estimation results.....	76
4.7.1 Estimated unknown parameters	78
4.7.2 Estimated passenger boarding and alighting flows, as well as passenger loads	83
4.7.3 Estimated passenger OD flows.....	87
4.8 Findings	91
4.9 Limitations and extensions	92
Chapter 5 Complex dwell time models	94
5.1 Introduction	94
5.2 Bus dwelling process	94
5.3 Information needed to understand the bus dwell time	98
5.4 Onboard survey in Kyoto City.....	99
5.4.1 Overview of the survey.....	99
5.4.2 Descriptive analysis.....	102
5.5 Modeling the interaction between buses and the effect of in-vehicle crowding.....	105
5.5.1 Modeling the effect of in-vehicle crowding	106
5.5.2 Modeling the interaction between buses.....	107
5.5.3 Regression analysis on the alighting time.....	108
5.5.4 Regression analysis on boarding time	109
5.6 Application of the complex dwell time model to OD estimation framework.....	110
5.6.1 Data detail level required and techniques for data processing.....	110

5.6.2 A case study in Kyoto City	111
5.7 Summary.....	115
Chapter 6 Conclusions.....	117
6.1 Summary of research	117
6.2 Contribution to existing knowledge.....	118
6.3 Future research directions	119
Reference	122

List of Tables

Table 2.1 Levels of spatial and temporal detail for data capture.....	12
Table 2.2 Functions and required data detail level.....	13
Table 2.3 Bus network planning	15
Table 2.4 Control strategies on bus bunching	23
Table 2.5 Bus arrival time prediction methods and data sources	25
Table 3.1 Coefficients of the independent variables in the LR model	40
Table 3.2 Coefficients of the independent variables in the LOGR model.....	45
Table 3.3 Four categories for binary classification results.....	46
Table 3.4 Performance comparison for 10-stop-ahead bunching prediction.....	51
Table 3.5 Supplementary information for ROC curves shown in Figure 3.8.....	53
Table 4.1 Regression analysis results of alighting time per passenger	75
Table 4.2 Elapsed time of each scenario	78
Table 4.3 Different prior information on pg and pa	80
Table 4.4 Prior information and posterior means of α , σW , σB and σA	82
Table 4.5 RMSE/MAPE of estimated passenger boarding/alighting flows and passenger loads	84
Table 4.6 Model fit and RMSE/MAPE of estimated passenger OD flows	89
Table 5.1 Bus route investigated in the onboard survey	100
Table 5.2 Data collected at the boarding door.....	101
Table 5.3 Data collected at the alighting door.....	101
Table 5.4 Regression analysis on the bus dwell time, a basic model	105
Table 5.5 Regression analysis on the crowding effect in bus dwell time.....	106
Table 5.6 Regression analysis on the effect of bunching and crowding in bus dwell time	107
Table 5.7 Regression analysis on alighting time WA	108
Table 5.8 Regression analysis on alighting time WA (without distinguishing methods of payment)	109
Table 5.9 Regression analysis on boarding time WB	109
Table 5.10 The simple and advanced dwell time models applied in demand estimation.....	112
Table 5.11 The prior knowledge on the stops of Kyoto City Bus No. 12	113
Table 5.12 RMSE and MAPE of the estimation results for each scenario.....	114

List of Figures

Figure 3.1 Data collected (left), data of Kyoto City Bus No. 205 (middle) and its configuration on real map (right)	38
Figure 3.2 Trajectories of Kyoto City Bus No. 205 in one day of April 2016	38
Figure 3.3 Headway fluctuation along the line for bunched buses	39
Figure 3.4 Performance comparison in terms of exact headway value.....	42
Figure 3.5 Performance comparison in terms of RMSE and MAPE under various prediction horizons.....	43
Figure 3.6 Performance comparison in terms of binary bunching identification.....	49
Figure 3.7 Performance comparison in terms of SES, SPC and ACC under various prediction horizons.....	50
Figure 3.8 ROC curves under various prediction horizons (1-stop, 5-stop, 10-stop and 15-stop ahead).....	53
Figure 4.1 Model framework; Items in grey boxes are the observations and data input; Items in white boxes are the unknown parameters and derived parameters	62
Figure 4.2 Patterns of boarding and alighting transactions observed from AFC data.....	75
Figure 4.3 Average alighting time against different number of alighting passengers	76
Figure 4.4 Posterior means of pg and pa	81
Figure 4.5 In-vehicle time distribution	82
Figure 4.6 Estimation results of boarding flows and OD flows by Scenario AT and T	83
Figure 4.7 Estimated passenger boarding flows	85
Figure 4.8 Estimated passenger alighting flows	86
Figure 4.9 Estimated passenger loads	87
Figure 4.10 Estimated passenger OD flows.....	90
Figure 4.11 Estimation errors of passenger OD flows by T^* ; Left: OD matrix errors highlighting the stop with largest errors, Right: Passenger flows destined to/leaving from the highlighted stop.....	91
Figure 5.1 Bus dwelling process – Dai et al. (2019).....	97
Figure 5.2 Event sequences from the viewpoint of bus m - Sun and Schmöcker (2018).....	98
Figure 5.3 Different sequences of passenger boarding and alighting flows observed from onboard survey	104

Figure 5.4 Estimated passenger boarding flows	114
Figure 5.5 Estimated passenger alighting flows	115
Figure 5.6 Estimated passenger loads	115

Chapter 1 Introduction

1.1 Background and research motivation

1.1.1 *Passive transit data*

Generally two types of passive transit data are distinguished (Trépanier and Yamamoto, 2015). The first one is voluntarily provided by the users, e.g. via using a smart phone application or carrying a portable GPS device. The second, and also the more common one, is collected without notifying the users, which includes the datasets from bus GPS devices, passenger counting sensors, smart card readers, roadside Bluetooth detectors or Wi-Fi package sensors, and so forth. Different from traditional active ways of data collection, such as travel survey which captures users' occasional travel behavior and in a small sample, passive datasets are in a significantly larger sample and can trace the dynamic travel behavior in a long term, albeit at a cost of missing socio-demographic attributes.

The emerging of passive transit data has drastically changed the ways of transit planning and operation, leading to a transition from traditional ways to data-driven methods. In planning aspects, data-driven bus route design and adjustment, optimized scheduling and timetable synchronization can efficiently accommodate the changing user demand including new transfer demand due to the expansion of the transit network. Furthermore, in contrast to static datasets such as survey data and map data, the availabilities of dynamic passenger and vehicle information provided by these passive transit datasets in particular offer advanced online operational strategies such as timed transfer scheme, bus signal priority, real-time bus headway control including bus holding, stop skipping and speed control, boarding limitation and bus injection. These control strategies enable the operators to maintain reliable service against the randomness due to passenger behavior, traffic flows, weather and so forth.

1.1.2 *The characteristics of AVL, APC and AFC data*

Automatic Vehicle Location (AVL) data provide a variety of possibilities to monitor, predict and control the service for the bus transit operators. AVL data usually contain 3-dimensional (latitude, longitude, time) coordinates which are recorded by a GPS device installed on the bus. The device restores the coordinate or uploads it to the server at a determined frequency. There is thus a time interval which is depending on the specification of the AVL system employed between two successive coordinates. The time interval could vary significantly (5 – 120 seconds) in different AVL systems. In some other cities, AVL data coordinates are more straightforward, presenting 2-dimensional

information (bus stop, arrival time) or (bus stop, departure time). Instead of a frequency, the coordinates are recorded by drivers' opening door or closing door or pressing a button.

Automatic Passenger Counting (APC) data is collected by the sensors installed at the bus doors. It provides reliable information on passenger boarding and alighting flows if they board and alight at separate doors in order. Operators can easily obtain 3-dimensional information (bus stop, number of boarding passengers, number of alighting passengers), and derive the number of onboard passengers by processing boarding-alighting data streams since the first stop. We note that the boarding and alighting flows are at aggregated level in APC datasets, and they provide valuable information on how busy a bus stop is as well as to where the system has capacity bottlenecks.

Automatic Fare Collection (AFC) systems require the user to tap a smart card at boarding and/or alighting in order to collect the fare. As a result, for each passenger, the boarding time, boarding stop, alighting time, alighting stop are recorded in the database, namely a time-stamped origin-destination (OD) entry. Here we assume that AFC data are available for a specific bus run. On one hand, it can replicate the AVL data to some degree. By taking the time stamps, it is feasible to reconstruct the 2-dimensional coordinate (bus stop, arrival/departure time) for the stops where passenger flows using smart card occur, although the reliability of inferred temporal information is strongly dependent on the number as well as the percentage of passengers using smart card. On the other hand, it can naturally reproduce the APC data by aggregating the passenger trips by bus stop. However, for some AFC systems e.g. a flat fare structure is employed, it may require the user to tap only at the boarding or alighting, then some extra efforts are needed to conduct the reconstructions of AVL and APC data, including destination inference, origin inference. Nevertheless, clearly AFC datasets are more informative than the other two ones, due to the unique disaggregated OD information and the capability of replicating the other two datasets

1.1.3 Understanding the pros and cons of AVL data

AVL systems are designed to monitor real-time location of the bus vehicle and investigate the actual deviation from the schedule. From the perspective of transit operators, lower cost is required by AVL systems than AFC and APC systems, since installing a small device such as a tiny chip for each vehicle can fully establish an AVL system thanks to the Global Position System (GPS). AVL systems can also include onboard computers that store event data such as stopping and starting to move, opening and

closing door (Furth, 2000). For the AVL systems collecting spatiotemporal coordinates as well as the event data, it is possible to obtain detailed time components in a bus trip including inter-stop travel time, dwell time, vehicle activity time and passenger activity time. If the event data record every stopping, the time spent on waiting due to traffic signals traffic congestion can also be produced by analyzing the stoppings that do not happen near the bus stops. Large quantities of observations gathered every day can indicate the mean, the confidence interval, the distribution of the travel time over a link and the dwell time at a bus stop, so that reasonable scheduling and travel time prediction become practical for the operators owning AVL data. We can conclude that the AVL system can provide accurate and plentiful information in terms of the location of the bus vehicles given a specific time point, or reversely the time given a location, e.g. arrival time given a bus stop. The noteworthy advantage of AVL data over APC and AFC is that the collection of spatiotemporal coordinates is independent of passenger flows so that AVL datasets contain more complete and reliable bus trajectories. In light of the continuous knowledge on the spatiotemporal movements of the buses, multiple-stop-ahead predictions on bus arrival times, bus headways and the bus trajectories have been investigated in many studies. The existing literature is reviewed in detail in Chapter 2.

The bottleneck of the dataset quality relies on the time interval of two coordinates or the data recording frequency. When the operators are trying to conduct a schedule adherence analysis or arrival time prediction, the information of concern or the needed model inputs are supposed to be the time points given the bus stops (exact arrival time or departure time at the stop is ideal). Long record interval may generate a bus trajectory consisting of sparse points whose spatial coordinate is not necessarily close to any of the bus stop, which can critically limit the usefulness of the AVL datasets. It may require extra effort to interpolate the data for the bus stops, and the interpolated time is likely to deviate from the truth. The difficulty of interpolation increases the longer the data interval between records. Furthermore, the data transmission sometimes is not reliable, due to signals being blocked by tall buildings, which further breaks the bus trajectory.

1.1.4 Bus bunching prediction: leveraging the strength of AVL data

With the help of bus AVL data, it is feasible to obtain the bus trajectories of successive buses, including the arrival time and departure time at each stop albeit interpolated time is of less reliability. These data can be used to predict bus bunching, and data-driven operation solutions. Bus bunching is the occurrence of irregular bus arrivals and uneven headways. It can be defined as the phenomenon of a

group of at least two successive services of a single bus line - or of multiple lines on a shared corridor - arriving at stops with a much shorter headways than the designed one. Bus bunching may be caused by the first service being delayed due to unforeseen traffic congestion along the route or unplanned high demand at previous stops. The subsequent service then has fewer passengers to pick up at that stop and departs earlier than scheduled. At downstream stops the effect is emphasized as the (small) delay to the first vehicle and the (slight) early arrival of the second vehicle result in increasingly longer dwell times for the first bus and increasingly shorter dwell times for the second bus. The time components such as dwell time, travel time, headway are available in the AVL datasets with some data processing, thus it is promising to analyse the causes and seek solutions for bus bunching problems based on AVL data. Bus bunching prediction is considered a useful application that leverages the data advantages.

1.1.5 Route demand estimation: bridging the gap between AVL data and APC/AFC data

An obvious shortcoming of bus AVL data compared with APC and AFC data is the lack of passenger demand information. Precisely speaking, AVL data is not designed to collect this kind of information. Ideally, combining bus AVL data with APC and AFC data can provide a full picture of the real problems that bothers the users, including delays, crowding, congestion and unreasonable route design. Thus it can help the operators to figure out better transit planning that ensures high punctuality/regularity and accommodates the demand dynamics better. A large body of literature integrate multiple datasets of AVL, APC and AFC to solve transport problems, which is reviewed in Chapter 2. It is undeniable that more limitations confront the operators that only have access to bus AVL data.

In order to support the “data-poor” operators, a primary objective of this dissertation is to explore a methodology that can successfully estimate passenger OD flows for a bus route using bus AVL data as the main data source. This might help to bridge the gap for the “data-poor” operators to the operators with access to more, and higher quality data. In this big data era, the “data-rich” operators utilize their data with advanced methodologies to improve their services further.

1.2 Research objectives

According to aforementioned background and research motivation, this research aims to achieve two main objectives using bus AVL data

1) Bus bunching prediction

Firstly, this research makes use of the rich information on time components in the bus trips and applies a logistic regression model to predict bus bunching events.

And through this analysis answer in particular following research questions.

- The identification of bunching event
- The selection of dependent and independent variable
- The difference between the newly proposed model and the existing literature that predict the bus headways, and more importantly, the advantages of the new method

2) Route demand estimation

Secondly, this research tries to explore the potentials beyond the design purpose of the bus AVL data, in order to make up the gap between AVL data and AFC/APC data. This research proposes a methodological framework to estimate OD flows for a bus transit route using bus AVL data. It cast the attention on the time components observed from AVL data such as dwell time and headway. It also investigates the estimation performance on passenger OD flows for each stop-level OD pair, for boarding, alighting flows and passenger loads (defined as onboard number of passengers at the arrival in this research) for each bus stop by aggregating the estimated OD flows.

And, as part of this, answer following research questions:

- The difference between expected OD rate matrix estimation and OD matrix reconstruction
- The effective methods to address the under-specification of this estimation problem given such limited and indirect observations on passenger flows
- Modeling the connection between the time components observed from bus AVL data and passenger flows
- The efficient solution algorithm
- The gap in estimation performance between using AVL data only and using additional AFC and APC
- How to maximize the improvement thanks to AFC and APC and meanwhile minimize the intervention degree.
- The reliability of the estimated results in terms of all the demand components: OD flows, boarding flows, alighting flows and passenger loads.

Also importantly, limitations and future extensions of estimating demand using AVL data is illustrated, with an emphasis on

- Other additional data sources that are easy to collect and can improve the estimation performance
- The feasibility of demand prediction in real time based on the proposed methodology

Finally, as a follow-up research objective of the second one, the effect of different dwell time models on estimation results is investigated. Complex dwell time models might be necessary for the cities where bus bunching and in-vehicle crowding are serious issues, since the relation between passenger flows and observed dwell time might greatly change. This research builds dwell times considering bus bunching and crowding, and tests the resulting estimation performance on demand.

1.3 Research contributions

Both research objectives are expected to significantly benefit the bus operators as well as extending the existing literature. The newly developed bus bunching prediction tool is expected to provide more reliable prediction results. The logistic regression estimates and predicts the probability of bunching occurrence, while existing methods predict headways and then judge the occurrence of bunching in a deterministic way. In long term or multiple-stop-ahead prediction, deterministic results are not convincing for the operators to make decision on controlling the bus in advance or not, as the prediction results inevitably tend to deteriorate.

The innovative methodology to estimate demand estimation from bus AVL data is considered to make noteworthy contributions in both methodological and practical aspects. It is a first attempt to solve such a highly underspecified problem as using observed time components to estimate passenger OD flows, which can be considered an important methodological contribution. In practical aspects, it explores the potential hidden in the bus AVL data in terms of inferring passenger boarding flows, alighting flows, passenger loads, crowding level, and OD flows. It illustrates the new roles that AVL data can play in improving bus transit service, which is expected to significantly help the operators that have difficulties in collecting AFC data and APC data.

This methodology parameterizes the OD flows by mean (expected) OD-paired arrival rate matrix in an observational period. Even if only bus AVL data observations are available, it can provide offline forecasts on the OD flows based on pre-determined schedule, online estimation on the boarding,

alighting and onboard number of passengers based on actual headway. It can be further applied to predict the boarding and alighting number of passengers, passenger loads and crowding level at the downstream stops, by combining the estimation model with dwell time/headway/bunching prediction models which have been already developed or newly proposed in this research.

1.4 Outline of the dissertation

After the introduction provided by Chapter 1, Chapter 2 reviews existing literature including a broad range of related aspects: 1) Studies using AVL data; 2) Studies combining AVL data with APC/AFC data; 3) Studies on bus bunching problem: corrective and predictive models; 4) Studies on OD estimation, in particular for public transport. Many of them employ AFC, APC and traditional survey data as the main data source, and some of them involve AVL data to enrich the model input or to solve the technical problems for which AVL data is more suitable. Then, 5) Studies on passenger load and crowding level estimation and prediction are reviewed.

Chapter 3 develops a bus bunching prediction methodology based on logistic regression and AVL data. It also introduces two existing methods using linear regression and support vector machine as the benchmark. The difference between the newly developed method and benchmark methods relies on that the former one directly computes the probability of a bunching event to occur while the latter ones compute the future headways and then judge if a bunching event is going to occur. A comprehensive comparison is elaborated to display the advantages of the new method. The data of a busy circular bus line in Kyoto City are used for the case study.

Chapter 4 proposes a methodology to estimate OD flows using bus AVL data as the main data source. A basic dwell time model is implemented to capture the connections between time components provided by AVL data and passenger flows. Considering the under-specification of this estimation problem, Bayesian inference is conducted to estimate the parameters and Markov Chain Monte Carlo (MCMC) methods are used as the solution algorithm. Estimation performance by using bus AVL data is compared with that by adding boarding or alighting counts data to the model input. The AVL data and AFC data are collected in a same observational period to validate the methodology.

Chapter 5 discusses complex dwell time models that consider the effect of in-vehicle crowding and bus bunching. In addition, it also investigates the improvements that are made by implementing

complex dwell time models on the demand estimation, compared with the performance in Chapter 4 based on a basic dwell time model. An onboard survey is conducted to collect the needed data in Kyoto City where busy and crowding bus services are challenging the operators.

Chapter 6 summarizes main findings of this research. Also it points out the limitations in the proposed bunching prediction tool and demand estimation methodology. Finally, it provides future research directions in light of the contributions of this research.

Chapter 2 Literature review

In Chapter 1, a brief introduction is conducted on the characteristics of three common passive transit datasets, in particular the bus AVL data. In this chapter, existing literature using one or multiple of these datasets is comprehensively reviewed, in order to form a better understanding on the roles that these datasets can play in solving transportation problems, and the state-of-art functions so far having been discovered for these datasets.

As is discussed in Chapter 1, AVL systems are effective at obtaining time components in the bus trips, and APC/AFC systems are powerful at inferring passenger flows. With a specific focus on the bus AVL data due to the reasons explained in the background, firstly this chapter aims to illustrate the advantages of bus AVL data and the state-of-the-art applications of bus AVL data. Secondly, it reviews the collaborations between the AVL data and APC/AFC data to improve the transit service, illustrating the limitation due to the lack of passenger demand data. Thirdly, it reviews the analytical and predictive control strategies that have been developed by a large body of literature towards bus bunching. In predictive control models, the AVL data and the collaboration among multiple datasets are playing crucial functions. It appears that the focus of the items to control is mainly narrowed on the headway, and it is difficult to control passenger flow if little knowledge concerning passenger demand is available, which cannot be inferred from the AVL data. Fourthly, the existing literature on OD estimation is reviewed. OD estimation is fundamental albeit challenging in demand estimation. Passenger boarding demand and alighting demand can be reliably derived if there is an inferred OD matrix for each transit route or the transit network. As a result, the expected dwell time and passenger load indicating capacity bottleneck at each stop can be estimated as well. Here the demand and dwell time estimation is considered equivalent to the offline demand planning. The online demand prediction is also developed by some studies with emphasis on passenger load and crowding level prediction

2.1 Overview of bus AVL data

Furth et al. (2003) define the detail levels of spatial and temporal information that can be captured by automated systems, mainly considering AVL and APC systems. Table 2.1 shows the data richness of various AVL-APC systems. Level A and B are both AVL systems of fairly low frequency, and Level B records the events such as stopping, moving or accident at sending the data to the server or local computer. Level C and D are the products of the AVL system collaborating with the APC system, so

that they record data for each stop instead of based on frequency. The data collection can be activated by door opening and de-activated by door closing, or triggered by driver's pressing a button to avoid missing the data for skipped stops. Level D records the events occurring during the inter-stop travel of the bus in addition to Level C. As the stop-focused bus operation data are of greater concern from the perspective of both passengers and operators, e.g. arrival time, crowding level, Level C and D are considered more useful than Level A and B, although they cannot provide the speed profile and indicate traffic condition along the bus route. Level E is continuously recording the data, and of the highest richness.

Furth et al. (2003) further summarize a variety of analysis purposes and decision-making objectives for the bus operators, and relate appropriate detail level to each objective. They suppose that the AVL system is to some degree tied with an APC system in Level C, D and E so that these levels also collect passenger demand information. We here present a modified discussion from their version based on our definition on those levels that passenger demand data is not available. We also change the function category. As is illustrated in Table 2.2, few analysis can be conducted with the very limited detail levels as A and B. They are only useful in obtaining rather general knowledge on the bus operation status, such whether the bus vehicle is currently on the route, whether any severe accident strikes the bus vehicle or the segment. It can roughly characterize spatiotemporal bus trajectories by interpolating, however the low frequency greatly reduces the reliability of the replicated trajectory. In fact, many bus AVL systems are between Level B and Level E, as they record the data at a frequency between every 5 seconds to 30 seconds. In these cases, it can generate a complete trajectory much closer to the truth than A and B, although the interpolated stop-based time information is still inferior to those provided by C, D and E. The bus AVL datasets we obtain from two Japanese cities: Kyoto and Shizuoka, for the case studies in Chapter 3 and Chapter 4 are of different detail level. In Kyoto, the data recording is planned to be conducted every 8 seconds but the frequency actually varies from 5 to 60 seconds perhaps due to some technical problems of data transmission. And it provides door opening or closing information for each recorded time point. Small movements over a series of time points with door closing can indicate congestion or signal effect given Geographic Information System (GIS) data. This dataset refers to a level between B and E, and it is inconclusive on its superiority or inferiority to C and D as it provides richer information on bus running over the segments albeit slightly less accurate with respect to stop-correlated time components. In Shizuoka, the operator only records the arrival time for each stop, which refers to Level C but underperforms some datasets of Level C that provide

as well departure time when it comes to the analysis requires dwell time and departure time.

We can conclude that Level B with a higher recoding frequency and Level C are capable of conducting a variety of analysis serving for operator's decision making. Precisely speaking, the desired detail level is the sources that can decompose the bus trip and obtain each time component such as bus arrival time, dwell time, departure time, segment running time and bus headway, and the noteworthy events or indicators of the events such as door opening/closing status and speed are a plus. The dwell time can be further decomposed into vehicle activity time and passenger activity time if the events indicators are available. With attaching passenger demand to the time point as bus arrival time or to the time period as bus headway, a primary 3-dimensional database integrating vehicle spatiotemporal information and passenger demand can be established, based on which almost all the analyses are feasible given other static data sources such as bus timetable and GIS data.

It should be noted that the demand analysis is currently impossible for the AVL system independent of APC or AFC system, at least according to the existing literature. Also it is difficult to investigate the real impact of delay and possible impact of schedule or route adjustment on the passengers with only the AVL data. The studies combining APC or AFC datasets with AVL data for demand analysis, estimation and forecast as well as bus control modeling are reviewed in Section 2.4 and 2.5. And a methodology using AVL system of Level C detail as the main data source to estimate demand is developed in Chapter 4 by this research.

Table 2.1 Levels of spatial and temporal detail for data capture

Level	Description	Event-independent records	Event records	Between-stop performance data
A	AVL without real-time tracking (or data transmission)	Infrequent (typically 60 to 120s)		
B	AVL with real-time tracking	Infrequent (typically 60 to 120s)	Each time point	
C	APC or Event recorder		Each stop	
D	Event recorder with between-stop summaries		Each stop and between-stop events	Recorded events and summaries
E	Event recorder/trip recorder	Very frequent (every second)	All types	All events, full speed profile

Source: Use of archived AVL-APC data to improve transit performance and management: Review and potential (Furth et al., 2003).

Table 2.2 Functions and required data detail level

Function	Analysis item	Least detailed level needed	External data needed
General performance analysis	Missed trips, off-route incidents	A	
	The reasons of incidents	B	
	Bus trajectory tracking and reconstruction	C or E	
Punctuality analysis	Schedule adherence	C	Schedule
	Impact of schedule deviation on passenger waiting time	C	Schedule
	The reasons of schedule deviation	D	Schedule
Run time analysis	Segment running time	C	
	Continuous running time	E	
	Impact of traffic signals	D or E	GIS, signal data
	Speed analysis	E	
Dwell time analysis	Delay due to unexpectedly long dwell time	C	
	Bus holding time	C	
Headway analysis	Stop-based headway	C	
	Continuous headway	A or E	
	Bus bunching	C	
	Impact of irregular headway on the passengers	C	APC
Demand analysis	Time-dependent OD demand	C	AFC
	Time-dependent boarding/alighting demand	C	APC
	Time-dependent load profile	C	APC
	Transfer demand	C	AFC
	Schedule rationality considering transfer demand	C	Schedule, AFC
Geographic and planning analysis	Mapping the bus trajectories with landmarks, road network and intersections	A	GIS
	Relating time-dependent passenger demand with geographic information	C	APC, GIS

2.2 Studies using bus AVL data as the main data source

Making use of the time component database obtained from the bus AVL data, two research objectives are mainly considered by the existing literature. Firstly, many studies measure the reliability metrics for the bus transit service of status quo, in order to understand the problems confronting the current service, such as service regularity indicated by headway and schedule adherence indicated by the deviation of arrival time from schedule. Provided with this correct knowledge on the inadequacies of the service, modifications can be applied to the schedule. However, offline data-driven planning and adjustment cannot guarantee that the service is always on schedule or of designed headway due to the randomness introduced by traffic flows and passenger flows to the system. Online operational strategies thus are required for maintaining reliable bus service. As a result, secondly some studies focus on prediction models on these time components such as arrival time and headway, which provides sources for real-time decision-making. Two types of equally important prediction should be distinguished in this problem, predicting the time point or location at which the unfavorable event might occur, and forecasting the consequences if control strategies are conducted.

2.2.1 *In bus network assessment and planning*

Ceder and Wilson (1986) broke down the comprehensive process of bus network planning into five levels: network design, setting frequencies, timetable development, bus scheduling and crew scheduling. A global review including 69 approaches that address the problems in the first three stages can be found in Guihaire and Hao (2008). They identify Stage 1 as strategic, and Stage 2 and 3 as tactical. The inputs and output of each stage are summarized in Table 2.3. Output of the previous stage together with the independent inputs frame the next stage planning. The foremost stage is to determine the configuration of a set of routes considering passenger demand and operator supply capacity of sustaining the current service and future extensions. Route performance indices can indicate the gap between demand and supply, also the need for new routes or route adjustment. Route performance indices can be obtained from the bus AVL data. In addition, the real arrival times, departure time and running times derived from the data suggest improvement directions on frequency and timetable for the bus planner.

Table 2.3 Bus network planning

Independent inputs	Planning activity	Output
Demand data Supply data Route performance indices	<u>Level A</u> Network design	Route changes New routes Operating strategies
Subsidy available Buses available Service policies Current patronage	<u>Level B</u> Setting frequencies	Service frequencies
Demand by time of day Times for first and last trips Running times	<u>Level C</u> Timetable development	Trip departure times Trip arrival times
Deadhead times Recovery times Schedule constraints Cost structure	<u>Level D</u> Bus scheduling	Bus schedules
Drive work rules Run cost structure	<u>Level E</u> Driver scheduling	Driver schedule

Source: Bus network design (Ceder and Wilson, 1986).

Chen et al. (2009) propose three indices to measure the reliability for transit bus at network, route and stop level, which are a punctuality index based on route (PIR), a deviation index based on stops (DIS) and an evenness index based on stops (EIS). PIR is the probability that a bus can arrive at the bus stops in a given time window. It is the mean value of the probabilities at all the stops of a route indicating a route-level schedule adherence degree. DIS then focuses on the stop-level reliability. DIS calculates the probability that the headway can be held within a threshold by every two successive buses at a specific stop. EIS is the headway evenness considering coefficient of variation of headways and it is especially useful to assess the service of high frequency. A large-scale on-board survey is conducted by more than 300 surveyors to collect the arrival times and departure times of 30 bus routes for 3 days in Beijing, China though, the time data can be gathered in a much more efficient way if bus AVL data is available. In order to calculate the reliability metrics weighted by boarding demand, they also collect the boarding numbers. Camus et al. (2005) investigate reliability using AVL data in Trieste, Italy. Their case study is based on a dataset of four routes during the peak hour of a month, which can characterize the daily fluctuating reliability performance. Lin et al. (2008) develop a quality control framework for bus schedule reliability using AVL data. They employ running time adherence and headway regularity as the key indicators of reliability, and their case study is based on 24 routes (48 route directions) over 29 weeks in Chicago, USA. In their AVL datasets, the arrival time of bus is recorded when it passes a series of locations whose number is fewer than that of bus stops on the bus route. These assessment on bus network reliability identify the need of schedule adjustment, and indicate the problem in network design if GIS data is integrated.

Bus AVL data also provide input for schedule planning algorithms. Bie et al. (2015) use bus AVL data to split the bus operating hours of one day into multiple time-of-day intervals characterized by observed bus dwell time and inter-stop travel time. They suggest that different schedule plans should be imposed on the buses in different operating interval. Bus dwell time is employed as an indicator of passenger boarding/alighting demand in various intervals, and inter-stop travel time indicates the traffic condition. Here bus dwell time is considered as a remedy for lack of directly observed passenger flows which can be provided by APC or AFC datasets. The case study is based on one bus route in Suzhou, China. The AVL data contains frequency-based spatiotemporal coordinates and speed of buses and the arrival/departure times are obtained by mapping the coordinates with GIS data. The data recording frequency is not disclosed though, the cutting of dwell time and travel time can be reliably

conducted with the speed information. As a more common function, the AVL data can provide observed stochasticity on bus arrival time, departure times, and running times, which can provide mean and confidence intervals for the schedule optimization algorithms. Mean values are useful for a variety of mathematical and heuristic approaches using deterministic assumptions on these time components (refer to the review by Guihaire and Hao, 2008). And confidence intervals can be the inputs to simulation approaches such as Bookbinder and Desilets (1992).

2.2.2 In bus operation

Bus AVL data has made noteworthy contributions to bus network assessment especially in terms of reliability analysis, and to schedule optimization though, the contribution is limited considering that bus network planning is user-oriented, the user-weighted reliability and the impact of schedule adjustment remain unmeasurable with the AVL data only. Nevertheless, it is rather more useful in bus operational aspects, such as predicting bus arrival times and running times. With this predicted information, a variety of control strategies have been developed to timely deal with a series of complicated operational problems due to the stochastic behavior of users and vehicles, and the resulting uncertainties. The real-time strategies have been designed for timed transfer (distinguished from schedule planning considering synchronization), bus bunching elimination or headway regularization and so forth. Recent research by Bie et al. (2020) comprehensively summarizes the strategies in the existing literature including bus holding strategy, signal priority, stop skipping and so on. Combination of multiple strategies are also reviewed. They review the literature from the perspective of bus bunching mitigation though, some of the control strategies are also widely used for timed transfer, such as bus holding (Dessouky et al., 2003), stop skipping (Nesheli and Ceder, 2014), short turn (Hadas and Ceder, 2010) and speed control (Hadas and Ceder, 2010). The main difference lies in the control objective.

Real-time control is unnecessary if bus arrival times and departure times are always as scheduled, which is almost impossible due to the randomness in inter-stop running times caused by uncontrollable traffic condition and in dwell times usually resulting from stochastic passenger behavior. In order to address the randomness issue in modelling a dynamic strategy albeit the lack of data, a lot of studies assume various probability distributions for these important time components as bus arrival time, travel time or delay from the schedule. Dessouky et al. (1999) summarize the distribution assumed by a number of studies. Normal, lognormal and Gamma distributions are commonly assumed for travel

time and arrival time, while exponential distribution is assumed for lateness and delay. Based on the random environment structured for bus operation, a simulation is then conducted to evaluate the control performance of the strategies (Andersson et al. 1979; Senevirante, 1990). The emerging of the bus AVL data enables these studies to calibrate the probability function with the observations in the real world, and to verify these methodologies with realistic settings. Dessouky et al. (1999) and Dessouky et al. (2003) propose real-time control schemes incorporating arrival time forecast model and dispatching/holding control model for timed transfer, the bus AVL data in Los Angeles, USA is applied in the simulation experiment. Nesheli and Ceder (2014) use the AVL data in Auckland, New Zealand and consider three real-time strategies: bus holding, individual stop skipping and segment (a series of successive stops) skipping for transfer synchronization. Similarly, bus bunching elimination models also require predicted time components in the bus trips to be the model inputs. Noteworthy studies by Moreira-Matias et al. (2016), Andres and Nair (2017), Berrebi et al. (2018) combine forecast and control models in bunching problem. The former two studies build forecast model on bus headway based on bus AVL data, and the third one forecasts the future bus trajectories using multiple data sources. The studies on the prediction models of bus trip time components using one or multiple data sources and the solutions on the bus bunching problem including analytical and data-driven control models are reviewed in detail in Section 2.3.

2.3 Studies combining bus AVL data and APC/AFC data

As is introduced in Chapter 1, the strength of the bus AVL data is richness in the time components of bus trips. The APC data can provide usually the start and the end of the passenger alighting/boarding process, which indicate the bus arrival time, bus departure time and bus dwell time excluding the vehicle activity time. The data quality is roughly the Level C detail. It can take the place of bus AVL data if Level C is enough for the analysis or modeling. Chen et al. (2009) collect the arrival time, departure time and boarding number at each bus stop to conduct passenger-weighted bus reliability analyses at stop, route and network level. APC datasets provide all the information needed for this analysis, while AVL data can only generate reliability indices without considering the weight of passengers. Bie et al. (2015) use bus dwell time as the indicator of demand to split the time-of-day interval. More reliable separation can be realized by directly using demand data if APC data is available. Cham (2006) build a reliability analysis framework based on AVL and APC data. The usefulness of APC data in explaining the causes of unreliability is illustrated. Strathman et al. (2003) investigate the causation between headway deviation and bus passenger loads by combining AVL and

APC data, and conclude that headway deviation is a primary cause to passenger overload.

The collaboration between AVL and APC/AFC data is very helpful for operational aspects. In many control strategy models, deterministic or probabilistic passenger boarding and alighting rates are assumed to generate passenger boarding and alighting flows. Real observations on passenger flows from the APC datasets can calibrate the incorporated passenger arrival and alighting models, and then make the experiment results more realistic and convincing. Berrebi et al. (2018) integrate the AVL data and APC data in Portland, USA as model inputs and conduct an experiment to compare the control performance of five bus control strategies previously proposed by Dagazo (2009), Xuan et al. (2011), Daganzo and Pilachowski (2011), Bartholdi and Eisenstein (2012) and Berrebi et al. (2015).

APC or AFC data is also used in prediction models on bus time components, independently or collaborating with the AVL data, and the related studies are reviewed in Section 2.4.2. The collaboration as well appears in OD estimation and passenger load estimation/prediction studies where the AVL data is no longer a main data source, and the related studies are respectively reviewed in Section 2.5 and 2.6.

2.4 Bus bunching problem

2.4.1 *Bus control strategies*

Most of the relevant existing literature can be cast into two categories according to their objective: bunching prediction and corrective strategies. A large body of literature discussed how to eliminate bus bunching using analytical or simulation methods following the seminal work by Newell and Potts (1964). Osuna and Newell (1972) and Newell (1974) tried to maintain the bus schedule by a single control point. In contrast, advanced control methods such as dynamic holding control proposed by Eberlein et al. (2001), Daganzo (2009), Xuan et al. (2011), Bartholdi and Eisenstein (2012), Zhang and Lo (2018) and velocity control developed by Daganzo and Pilachowski (2011). Stop skipping discussed by Sun and Hickman (2005) also assumes frequent and efficient communication between bus drivers and the control center. Berrebi et al. (2018) tested the control strategies proposed by Dagazo (2009), Xuan et al. (2011), Bartholdi and Eisenstein (2012), Daganzo and Pilachowski (2011) on a bus route in Portland, Oregon. The experiment was based on real bus AVL (Automatic Vehicle Location), APC (Automatic Passenger Counter) and traffic signal data. The effectiveness of each strategy to stabilize bus headways was confirmed. Further, the effect of incorrect future headway

prediction on each strategy was discussed. The variance of controlled headway was found rising significantly as the prediction errors increased. Instead of actively adjusting the headway, Schmöcker et al. (2016), Wu et al. (2017), Sun and Schmöcker (2018) discussed passenger re-distribution and overtaking strategies which are activated when bunching occurs. These strategies aim to equalize passenger boarding numbers for bunched buses through queue management.

Here we classify the existing literature according to the type of control strategy, with a focus on the data required for model validation and whether a prediction module is incorporate to generate the model inputs. We can summarize mainly eight types of bus control strategies in addressing the bus bunching problem: bus holding, bus signal priority, speed control, stop-skipping, short turn, overtaking, boarding limitation and bus injection.

Bus holding is the most commonly considered strategy. It extends the dwell time of a bus at a control point, usually an intermediate stop or terminal stop. We can simply distinguish three sub-categories for bus holding: based on schedule, based on forward-looking headway, and based on backward-looking or two-way looking headway. Holding based on schedule is usually conducted at one or multiple pre-determined controlling stop. The dwell time is forced to extend so as to maintain the schedule. Schedule adherence and minimizing passenger waiting time are the objectives. However, it is likely to greatly increase the total time of a bus trip and reduce the bus frequency. Besides, it is not effective to correct the delayed bus and improve the service regularity. Daganzo (2009) proposes a forward-headway-based approach to dynamically control the bus using a headway threshold, the holding time is a function of scheduled headway, actual headway and adjustment parameter. We note that prediction is not required in this case where only actual forward headway is the dynamic model input. At the arriving of the bus of concern, the forward headway from its previous bus is always measurable. This method requires few data and is easy to implement. It significantly increases the bus travel speed comparing with the schedule-based strategy. Bartholdi and Eisenstein (2012) suggest the importance of considering backward headway, and Daganzo and Pilachowski (2011) propose a combined strategy of speed adjustment and holding based on two-way headway. The calculation of backward headway involves a prediction on the arrival time of the next bus. This kind of prediction becomes possible with the emerging of bus AVL data. And reliable prediction can be obtained as the prediction horizon is usually short, considering the future arrival time of one or two buses to the control stop. Berrebi et al. (2015) consider a horizon of n buses coming to the control stop, and the backward

headway is calculated based on the joint probability distribution of next n bus arrival times.

Most of the studies on control strategies employ a bus propagation model to simulate the outcomes. Forecast on resulting headway, or passenger waiting time determines whether the control is to be conducted. AVL and APC data are interpreted as historical observations from which the distribution of link travel time and passenger boarding/alighting demand can be inferred. Some studies draw the travel time or demand from the distribution randomly to replicate the stochastics in reality, while some others use mean values to simplify the bus propagation model. Another main stream is noting the function of data and assuming virtual constant/random travel time/passenger demand (Berrebi et al., 2015; Sánchez-Martínez et al. 2016). In Table 2.4, AVL or APC in bold denote that the real data is applied to obtain the distribution or mean (Wu et al., 2017; Bie et al., 2020), otherwise the data is simulated in the studies. It obviously shows that AVL data is widely used or assumed by the studies. APC data is used or assumed by the studies paying specific attention on passenger demand, such as boarding limitation and stop-skipping.

Many studies model the passenger boarding and alighting demand independently. They assume passenger arrival rate/boarding rate for each stop, based on APC data (Wu et al., 2017) or purely assumption (Berrebi et al., 2015). For the alighting demand, Liu et al. (2013), Berrebi et al. (2015), Chen et al. (2015) and Wu et al. (2017) assume that the passenger boarding at a specific stop will evenly alight at the remaining stops. In this way, passenger arrival rate and alighting probability of all the on-board passengers can be obtained for each bus stop. Delgado et al. (2012), Schmöcker et al. (2016), Sun and Schmöcker (2018) ignore the alighting process in their models, which further simplify the randomness of passenger behavior. Sánchez-Martínez et al. (2016) is the only research that models the demand with time-varying mean arrival rates at the origin-destination pair level, assuming the availability of time-dependent origin-destination matrices or AFC data. They note the recent research in the use of automatically collected data to infer travel patterns enables estimating dynamic origin-destination matrices (Gordon et al., 2013). Matrix estimation methods such as iterative proportional fitting may be used when only boarding and alighting data are available (McCord et al., 2010). This OD-pair arrival rate assumption is most realistic but requires highest quality of data which is only available in onboard survey or AFC data. Boarding limitation, stop-skipping and overtaking are the strategies that control or affect passenger boarding behavior, however the simplified alighting behavior might make the controlling results biased. The considerations on the connections between passenger

boarding and alighting behavior e.g. OD-pair arrival rate, are expected to uncover new insight on the strategy performance. OD-wise assumptions make the random alighting demand predictable. Limiting the boarding passenger at a specific stop may consequently limit the alighting at another specific stop, if dominant OD-pairs exist. New strategies can be developed to split and limit the boarding passenger to not only alleviate in-vehicle crowding, but also create more stop-skipping opportunities due to the alighting demand at the skipped stop has been already controlled. Furthermore, AFC data or OD information is very helpful to establish effective short turn strategies as in Cortés et al. (2011) and Tirachini et al. (2011).

Table 2.4 Control strategies on bus bunching

Control strategy	Author(s)	Data for model validation	Prediction for model inputs
Bus holding, based on	Schedule	Osuna and Newell (1972)	AVL, APC
		Newell (1974)	AVL, APC
		Xuan et al. (2011)	AVL
	Forward	Eberlein et al. (2001)	AVL
	headway	Daganzo (2009)	AVL
	Backward	Bartholdi and Eisenstein (2012)	AVL
	or two-way headway	Berrebi et al. (2015)	AVL, APC
		Sánchez-Martínez et al. (2016)	AVL, AFC
		Andres and Nair (2017)	AVL
		Zhang and Lo (2018)	AVL
Bus signal priority	He et al. (2016)	AVL (bus and traffic), APC	Headway, passenger load, traffic flow
Bus signal priority and holding	Koehler and Kraus (2010)	AVL	Traffic flow
	Estrada et al. (2016)	AVL	Headway
Speed control and holding	Daganzo and Pilachowski (2011)	AVL	Headway
Speed control and signal adjustment	Bie et al. (2020)	AVL, APC	Travel time
Stop skipping	Sun and Hickman (2005)	AVL, APC	No
	Liu et al. (2013)	AVL, APC	No
	Chen et al. (2015)	AVL, APC	No
Stop skipping and holding	Moreira-Matias et al. (2016)	AVL, APC	Travel time, headway, bunching likelihood
Short turn	Cortés et al. (2011)	OD, AFC	No
	Tirachini et al. (2011)	OD, AFC	No
Overtaking	Schmöcker et al. (2016)	APC	No
	Sun and Schmöcker (2018)	APC	No
Overtaking and holding	Wu et al. (2017)	AVL, APC	Headway

Control strategy	Author(s)	Data for model validation	Prediction for model inputs
Boarding limitation and holding	Delgado et al. (2009)	AVL, APC	Headway, demand
	Delgado et al. (2012)	AVL, APC	Headway, demand
Bus injection	Morales et al. (2019)	AVL	Headway, demand

2.4.2 Bunching prediction

Rather than predicting bus bunching events, most existing literature focuses on bus arrival time and headway. Though clearly closely related, this literature can again be grouped into three subcategories: bus trajectory, bus arrival time and headway prediction. Complete bus trajectory prediction is most challenging but also most informative. It provides predicted stop arrival and departure time, stop-to-stop travel time, as well as the headway between consecutive buses for bus operators and users. Hans et al. (2015) developed a sequential mesoscopic simulation that elaborately considered the stochastics generated during bus dwell time and link travel time. A bundle of possible future trajectories is simulated based on the distribution assumed for the time components in a bus trip and the associated parameters are calibrated with AVL, APC and traffic signal data. This method delivering robust prediction results to the operator. Distribution or range for future arrival time and headway can also be easily obtained. A shortcoming of this method is that the predicted range of arrival time or headway might be too wide to be conclusive for operators' decision making. Recent research by Dai et al. (2019) also modeled bus dwell time and link travel time in detail to reproduce the trip travel time variability for a bus line. They specifically considered the bus waiting time due to the interaction between buses at the stop intersected by multiple bus lines, which is also defined as common-line bunching in Schmöcker et al. (2016). They inferred the probabilities of the bus from a specific line queueing (bunching) after the other common lines at the stop from bus GPS data. Yu et al. (2016) and Yu et al. (2017) conducted a solid literature review on the methods addressing bus arrival time/travel time prediction. They reviewed the implemented data source and algorithm of each relevant literature. Here we re-classify the reviewed literature and add some new studies in Table 2.4. It shows that SVM (Support Vector Machine), KF (Kalman Filter), KNN (K-Nearest Neighbor), ANN (Artificial Neural Network) and regression-based methods are frequently used. Yu et al. (2011) used SVM, ANN, KNN, and LR to predict arrival time for a 0.7km common line section where more than 10 bus routes overlapped in Hong Kong. Kumar et al. (2018) combined KF and KNN to tackle the prediction of bus travel time and arrival time. In this hybrid model, KNN classifier is used to refine the model input of

KF model.

Future headway is the difference between the predicted arrival times of two consecutive buses and can be obtained by arrival time prediction methods. There are also some studies directly focusing on the prediction of headway itself. Yu et al. (2017) proposed a probabilistic prediction approach using RVM (Relevance Vector Machine) to attach a confidence interval for each predicted headway for 2- and 3-stop-ahead. RVM presents better robustness by comparing the results with the deterministic single values derived by SVM, KF, KNN and ANN algorithms. Andres and Nair (2017) integrated headway prediction and bus holding control strategies. Regression, ANN and autoregressive models are used in their work to predict future headways with 5min and 10min prediction horizons. The prediction results are applied as input to an analytical model extending Daganzo (2009).

Table 2.5 Bus arrival time prediction methods and data sources

Author(s)	Data	Algorithm
Chien et al. (2002)	AVL (historical) data	ANN
Jeong and Rilett (2004)	AVL data	ANN
Yu et al. (2006)	AVL and weather data	SVM
Chang et al. (2007)	AVL data	KNN
Coffey et al. (2011)	AVL data	KNN
Yu et al. (2011)	AVL data	SVM, ANN, KNN, LR
Liu et al. (2012)	AVL data	KNN
Sinn et al. (2012)	AVL data	Kernel regression
Moreira-Matias et al. (2016)	AVL and historical data	ANN
Chen et al. (2004)	APC data	ANN and KF
Patnaik et al. (2004)	APC data	Regression
Shalaby and Farhan (2003)	AVL and APC data	KF
Lin et al. (2013)	AVL and AFC data	ANN

Although headway prediction methods have made great advancement, it remains a challenging work to successfully identify coming bunching events in multiple-stop-ahead prediction. The accuracy of

bunching prediction is heavily dependent on the reliability of headway prediction whose results deteriorate gradually as the prediction horizon extends. Yu et al. (2016) used several well-developed algorithms to predict headway first then convert the result to binary bunching occurrence. RMSE for headway prediction is 2min and 99% sensitivity is realized for bunching in 2-stop-ahead prediction, but the performance deteriorates to 6min for RMSE and 73% for sensitivity for 5-stop-ahead prediction. Moreira-Matias et al. (2016) built a regression-based model to predict the headway for a downstream stop and calculate the likelihood of bus bunching to occur for all the further downstream stops. The focus of their study was to propose a proactive control framework in which every suspicious event triggers a bunching alarm. The effect of bunching likelihood thresholds was not investigated. It should be noted that Moreira-Matias et al. (2016), Andres and Nair (2017), Berrebi et al. (2018) combined prediction and correction, and tested the feasibility and benefit of putting corrective strategies into practice. Instead of bunching prediction, Arriagada et al. (2019) used bus GPS data and smartcard data to investigate the causes of bus bunching, with an emphasis on the planning side. Scheduled frequency, stop location and configuration (number of the berths), traffic signal and bus lane design are found to be influential. Such research provides insight into bunching prevention in the planning stage.

It can be concluded that few literature considers a straightforward method to predict bus bunching. Headway-based prediction method for bus bunching is effective as many strategies are based on headway, such as bus holding. Some strategies instead are triggered by bunching events or other indicators, such as overtaking and boarding limitation. We here aim to extend the family of prediction tools for bus bunching by providing a new direct method to predict bunching event.

In addition, few literature models passenger demand at OD-pair level in the bus propagation model for bus bunching, which reduces the contribution to real operation. We suppose the simplification on passenger demand also results from the difficulty to obtain OD data. Obtaining APC data appears easier and we will discuss the method to estimate OD matrix from APC data in the next section. Table 2.4 also shows that AVL data is more widely used in the studies concerning bus bunching. We expect more strategies to be developed if the OD matrix can be inferred from bus AVL data, which further enhances the contribution of this research.

2.5 OD estimation

In this section, a comprehensive literature review on OD estimation is provided to overview the contributions and challenges of estimating OD matrices from bus AVL data. The OD estimation problem has received long-lasting attention in transportation studies as it plays a fundamental role in transport planning and management. Clearly, the OD flow estimation performance strongly depends on the quality of the observations. It is a highly underspecified problem as in most cases the number of observations is much less than the number of OD pairs in the network. Early seminal studies estimate traffic OD flows from link counts and employ a base OD matrix as prior information to counter the underspecification problem. Estimation methods include the maximum entropy and minimum information principles (Van Zuylen and Willumsen, 1980), Bayesian inference (Maher, 1983), maximum likelihood models (Spiess, 1987) and generalized least square approaches as in Bell (1991). The studies related to the latter three methods are also reviewed by Cascetta and Nguyen (1988).

For the estimation of public transport OD flows in urban networks similar approaches can be used although an additional problem is the need to infer transfer flows. Onboard OD surveys are a straightforward way, but need to be conducted manually and are hence infrequent, randomly capturing travel behavior from a small sample size of transit users. Unchained boarding-alighting count data which can be massively recorded by APC devices are considered as an alternative. Ben-Akiva (1985), Mishalani et al. (2011) and Ji et al. (2015) combine unchained boarding-alighting counts and a base OD matrix derived from onboard survey data to update OD flows, based on the Iterative Proportional Fitting method. Mishalani et al. (2011) find that increasing the sample size of the onboard OD survey improves the estimation performance. Li and Cassidy (2007) and Li (2009) estimate the OD flows for a transit route using the count data only. It is an underspecified problem as the observed counts are only two vectors characterized by the number of the stop. Inferring connections between the vectors are required to transform the independent boarding and alighting flows into a matrix. Li and Cassidy (2007) introduce the conditional probabilities of a passenger alighting at a specific stop given a boarding stop as parameters for the OD matrix. They categorize all the bus stops into major and minor stops to greatly decrease the number of parameters. The parameters are estimated by balancing methods using counts data and calculated passenger loads. Li (2009) reduces the number of parameters for the alighting probability matrix by a Markov chain model. The probability of a passenger's alighting at a specific stop is considered only dependent on whether he/she is on-board at the previous

stop so that the alighting probability matrix is reduced to a vector. Hazelton (2010) suggests that this is unrealistic and proposes an MCMC method interpreting the transition probabilities derived by Li (2009)’s model as the proposal distribution in Metropolis sampling.

The emergence of smart cards and other electronic payment systems has drastically changed the status of the OD estimation problem for transit routes. If the transit system requires the passenger to tap a smart card both when boarding and alighting, the transaction data can be naturally feed into a complete OD matrix. However, the boarding or alighting entries might be missing if only one tap is required, e.g. a flat fare structure is applied on the transit route. Barry et al. (2002), Trépanier et al. (2007) and Zhao et al. (2007) discuss the approaches of destination inference and OD matrix estimation using smart card datasets without alighting entries. A unique user ID is recorded in smart card datasets based on which personal trip itinerary can be created. By applying distance and time thresholds, the alighting stop can be inferred. Munizaga and Palma (2012), Gordon et al. (2013) use AFC and AVL data to infer the origin-destination flows of AFC users including their transfer points. They also discuss expansion methods to infer the flows of all users, i.e. including those who do not use the smart card for payment. Gordon et al. (2018) investigate the estimation of the “origin-interchange-destination” flow problem containing both AFC and non-AFC users by combining APC data with AFC and AVL data. Detailed reviews on the studies using smart cards can be found in Pelletier et al (2011) and Hickman (2017). Pelletier et al (2011) review the studies regarding transit planning and categorized them into strategic-level, tactical-level and operational-level. Hickman (2017) narrows the focus on transit OD estimation.

However, as noted in the introduction, transit operators do not always have access to smart card or passenger count data. Cash payment is still prevalent in many cities. In other cases passengers might not have to pay at all, particularly commuters or concessionary passengers are often only showing their travel permission at boarding or alighting without leaving any record of having been on-board. Therefore, we consider the problem of estimating OD flows given very few observations on passenger flows. We focus on using bus AVL data in this thesis and seek the solution from Bayesian inference which is proved powerful in inferring OD trips given limited observations or observations with uncertainty by Maher (1983), Hazelton (2001), Hazelton (2008) and Hazelton (2010). Such Bayesian inference frameworks are found useful also in solving other transportation problems, e.g. using travel time observations to estimate the link cost and the route choice parameters simultaneously in a metro passenger assignment problem by Sun et al. (2015), and missing traffic data imputation by Chen et al.

(2019). The posterior distribution of each parameter derived by Bayesian inference can provide the transit operators with the confidence interval of the estimated OD flows, which better accounts for the inherent randomness existing in passengers' behavior. The main challenge of this study is that we try to estimate OD flows by taking a series of bus dwell times obtained from bus AVL data as the observations, which makes it methodologically different from the existing literature. The problem we tackle is more underspecified in that unique solution is neither available for boarding and alighting number given dwell time, nor for OD flows given unchained boarding-alighting flows.

Hazelton (2001) emphasizes the subtle distinction between two problems in OD inference: reconstructing the actual number of OD trips and estimating mean OD trip rates (expectation) which are also defined as the unknown parameters in contributions such as Spiess (1987); Li and Cassidy (2007) or Li (2009). The aim of reconstruction is to precisely replicate the actual number of OD trips during the observational period while that of estimation is to infer the expected number of OD trips. Here, we also distinguish the reconstruction problem whose objective is to accurately estimate the OD flows for each bus run in an observational period and the estimation problem which is to infer the mean (expected) OD flows per bus run in that period. We parameterize the OD matrix by a passenger arrival rate matrix. The alighting probability matrix assumed by Li and Cassidy (2007) and Li (2009) is not suitable for our problem in that passenger counts are not observable. We put the focus on the mean estimation problem in this thesis, as the reconstruction can be considered the next step to parameter estimation, which is also pointed out by Li (2009).

2.6 Passenger load estimation and prediction

Li and Hensher (2011) find that one minute of standing travel time is 2.04 times as one minute of traveling with a seat, and one minute of waiting on a crowded platform is 1.7-2.5 times as that of waiting on a normal platform. In order to deliver comfortable transit service for the users, crowding problem should be seriously addressed. Passenger load estimation and capacity bottleneck analysis are important in demand planning and the prediction is crucial in the operation. Provided with the OD matrix, the mean number of boarding and alighting passenger can be naturally obtained, the mean passenger load can also be estimated by tracking the expected boarding and alighting dynamics from the first stop. The mean passenger load evolution over the bus route can indicate the capacity bottleneck and crowding level profile. Frequency and schedule adjustment can be applied to address the possible crowding in the planning stage, short turn and demand limitation are then operational

strategies. Zhu et al. (2017) propose a passenger-to-train assignment model for railway transit using AVL and AFC data. Given the boarding and alighting station obtained from AFC data, they propose a probabilistic model to assign the passenger to the possible trains based on the train trajectories which are inferred from AVL data, and walking time distribution for access time (from the gate to the platform) and egress time (from the platform to the gate). They also consider the possibility of boarding failure due to crowding. Jenelius (2019) predicts the vehicle-specific (one metro train contains multiple vehicle) passenger load and crowding level for metro transit in real time using train load data collected by air suspension system on the vehicle. So far, only few literature discuss the prediction on bus passenger load in real time. Zhang et al. (2017) combines AVL data and AFC data to predict passenger flows for bus transit. Historical OD data in archived AFC data is used to learn passenger alighting pattern given a specific boarding stop. The prediction is then realized given real-time bus AVL data and boarding data. We consider the inferred mean arrival rate at OD-pair level from bus AVL data in this research can also contribute to the literature on prediction.

Chapter 3 **Bus bunching prediction**

3.1 Introduction

Bus bunching is a frequently occurring undesired event. Generally it can be defined as the phenomenon of two successive bus runs of a single line arriving at a stop within significantly shorter headways than the designed one. Bunching involving more than two buses is also regularly observed. Bus bunching may be initiated by the arrival of one bus run being delayed at an upstream stop. More passengers are likely to accumulate for the delayed bus at that stop and the bus is thus further delayed. Conversely, the subsequent run has fewer passengers to pick up and departs earlier than scheduled. Accumulated delay to the first vehicle and increasingly earlier arrival of the second one result in obvious inequality in dwell times and on-board passenger numbers. As the inequality aggravates over a sequence of stops, the scheduled headway is significantly shortened or eventually offset and the leading bus among bunched bus is often overcrowded.

Accurate prediction on headway or bunching itself can help to spotlight the coming bunching and further assist the operator to eliminate bunching in real time. A useful prediction tool is expected to a) have a long enough prediction horizon to allow the operator's implementation of countermeasures and b) provide information on the reliability of the prediction. The latter point is important in order to account for different preferences among operators. A bunching-averse operator is willing to frequently control the service to avoid any possible bunching, whereas some other operators may hesitate to take control action that will negatively impact some passengers, they thus only correct the predicted bunching of high confidence level. Therefore, this thesis suggests a probabilistic binary prediction method.

This chapter aims to extend the existing literature in two aspects. Firstly, it builds a LOGR (Logistic Regression) model to predict the likelihood of bunching to occur using bus GPS data, and tests the prediction performance under a wide range of prediction horizons varying from 1-stop-ahead to 15-stop-ahead, with an emphasis on multi-stop-ahead prediction and understanding the regularity deterioration pattern. Secondly, this study tries to enhance the robustness and flexibility for existing prediction tools. To achieve this ROC (Receiver Operator Characteristic) curves are utilized. This method is widely used in evaluating the performance of binary classification models and in this study it is interpreted as the optimal front of the proposed LOGR. This chapter also explains how to conduct

the trade-off between “sensitivity” and “specificity” from an operator’s perspective.

The Chapter is organized as follows. After this introduction, the predictive methodology using logistic regression is elaborated in Section 3.2. We point out that logistic regression might be biased when used for “rare events data” as is the case in our example and provide a correction method. Then two headway-predicting algorithms: LR (Linear Regression) and SVM (Support Vector Machine) are taken as the two benchmark approaches in this study and are also briefly introduced in this section. In Section 3.3, the characteristics of the collected data are described, including data collection period, average stop-to-stop travel time, average scheduled headway, fluctuation patterns for headway, etc. Based on this, appropriate prediction horizons and bunching thresholds are determined. The case study is described in Sections 3.4-6. In Section 3.4, prediction performance of the two headway-predicting algorithms is discussed. The prediction performance of the proposed LOGR is evaluated and compared with headway-based methods in Section 3.5. The trade-off functionality of LOGR is discussed in Section 3.6. Conclusions and further work can be found in Section 3.7.

3.2 Bus bunching prediction

3.2.1 The identification of bus bunching event

As a bunching event involves two buses we refer to these as front bus and back bus respectively. Let a binary variable b_m^n denote whether bus run m is caught in bunching as the back bus during its dwelling at stop n . a_m^n and d_m^n denote the arrival and departure time of bus run m at stop n respectively. At stop n , for each bus run m ($m \geq 2$) we can obtain $\Delta_{m-1,m}^n$ which is the time interval between the arrival time of bus m and the departure time of bus $m-1$ in Eq. (3.1). Bus run m is considered bunched with bus run $m-1$ at the stop when $\Delta_{m-1,m}^n$ is below a threshold Δ_0 . The threshold can be determined by the operator. Yu et al. (2016) and Moreira-Matias et al. (2016) used 1/4 of the scheduled headway. $\Delta_{m-1,m}^n$ is defined as the departure-to-arrival headway in this study. Different from arrival-to-arrival or departure-to-departure headway, $\Delta_{m-1,m}^n$ is negative when two buses overlap at the stop. As overtaking is not allowed, for each stop n , bus $m-1$ always arrives and departs earlier than bus m , and accordingly time interval $\Delta_{m-1,m}^n$ can always be obtained before the departure of bus m .

$$\Delta_{m-1,m}^n = a_m^n - d_{m-1}^n \quad (3.1)$$

For each bus m ($m \geq 2$), the binary bunching status b_m^n can be derived by Eq. (3.2)

$$b_m^n = \begin{cases} 1, \Delta_{m-1,m}^n \leq \Delta_0 \\ 0, \Delta_{m-1,m}^n > \Delta_0 \end{cases} \quad (3.2)$$

3.2.2 Variable selection

Following afore reviewed literature the continuous $\Delta_{m-1,m}^n$ can be used as the dependent variable for headway-prediction approaches. For bunching prediction then an additional step is required judging whether the predicted headway is below a prior defined bunching threshold or not. Instead, in this study, b_m^n is used as dependent variable using logistic regression to directly predict the binary bunching status and bunching probabilities.

Gradually accumulated or suddenly significant inequality in dwell time and travel time might lead two successive buses to be bunched. The back bus in a bunching event tends to have a shorter forward-looking headway, negative deviation from timetable (ahead of schedule), less on-board passengers and shorter dwell time than those of front buses in a bunching event or of non-bunched buses (Degeler et al., 2018). Yu et al. (2016) used boarding and alighting numbers of two successive buses, link travel time and headway at an upstream stop as the input to their headway-based prediction approach. As only bus GPS data is used in this study, information regarding boarding, alighting as well as on-board passengers are not available. Instead dwell time is included in the variable set in addition to headway. Deviation from the timetable is excluded here, as bus dispatching is not based on the timetable in some cities and the data for this variable might not be available. To conclude, dwell time of two successive buses and their headway at an upstream stop $n-k$ are used as the main leading indicators of a coming bunching event in the k -step-ahead prediction. The detailed notation is as follows:

t_m^{n-k}	dwell time of bus run m at stop $n-k$
t_{m-1}^{n-k}	dwell time of bus run $m-1$ at stop $n-k$
$\Delta_{m-1,m}^{n-k}$	time interval between the arrival time of bus m and the departure time of bus $m-1$ at stop $n-k$
k	prediction horizon in terms of number of stops, $k = 1, 2, 3, \dots, N-1$ and N denote the last stop of the bus route

We always have $n > k$, so that this method cannot conduct the k -stop-ahead prediction until the bus run m passes the initial k bus stops, e.g. The prediction starts from stop 6 in the 5-stop-ahead prediction, using the data at stop 1. Also note that $m \geq 2$ and the first bus has zero probability to be bunched as the back bus.

3.2.3 Logistic regression

LOGR (Logistic Regression) modeling is widely used in classification problems. In binary classification it not only helps to categorize observations into positive or negative classes, but also interprets the causality by producing the significance of each independent variable. Moreover, it computes the probability of each observation to be in the positive or negative class. The binary bunching status from the perspective of the back bus b_m^n ($m \geq 2$) is taken as the dependent variable. t_m^{n-k} , t_{m-1}^{n-k} , and $\Delta_{m-1,m}^{n-k}$ are the independent variables. Let $\mathbf{X}_m^n = [t_m^{n-k}, t_{m-1}^{n-k}, \Delta_{m-1,m}^{n-k}]$, then probability of bus run m being bunched at stop n as a back bus can be derived as

$$Pr(b_m^n = 1 | \mathbf{X}_m^n) = \frac{1}{1 + e^{-\boldsymbol{\beta} \mathbf{X}_m^n}} \quad (3.3)$$

with parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]$ obtained by fitting the model with real data. Then $Pr(b_m^n = 1 | \mathbf{X}_m^n)$ for each bus run m at any stop n for the same or a different data sample can be computed. b_m^n is predicted to be positive (one-event) if $Pr(b_m^n = 1 | \mathbf{X}_m^n)$ exceeds a probability threshold Pr_x which is also known as the cut-off point, otherwise, negative (zero-event).

$$b_m^n = \begin{cases} 1, & Pr(b_m^n = 1 | \mathbf{X}_m^n) > Pr_x \\ 0, & Pr(b_m^n = 1 | \mathbf{X}_m^n) \leq Pr_x \end{cases} \quad (3.4)$$

3.2.4 Rare events bias

Irregular arrivals are common in bus transit operation, however few of them turn into severe bunching. The bunching probability which is the ratio of bunching occurrence to the total number of dwelling in Yu et al. (2016) varies from 3% to 17%, from 0.15% to 7.17% in Moreira-Matias et al. (2016), and from 3% to 9% in our 5-day testing data. Bunching is hence a “rare” event in the dataset.

“Rare events data” refer to large datasets in which it is significantly less likely that the binary

dependent variables take one than zero. King and Zeng (2001) considered events such as wars, natural disasters or epidemiological infections within long term time series data. They found logistic regression underestimates the probability of rare events because they tend to be biased towards the majority class, which is the less important class in most cases. This can be explained as follows:

The dependent variable Y_i follows a Bernoulli probability distribution that can take the values of one and zero with probabilities π_i and $1 - \pi_i$ respectively. The probability function can be written as

$$Pr(Y_i|\pi_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i} \quad (3.5)$$

It is easy to derive the expectation and variance of Y_i as

$$E(Y_i) = \pi_i \quad (3.6)$$

$$V(Y_i) = \pi_i(1 - \pi_i) \quad (3.7)$$

If the regression model has some explanatory power, the variance in the dependent variable has to be large enough. The variance becomes larger as π_i increases and reaches its maximum if $\pi_i = 0.5$, which indicates that it is favorable to involve an equal number of ones and zeros in the dataset. Cosslett (1981) and Imbens (1992) also showed that equally sampling the two classes is optimal.

King and Zeng (2001) further discuss that selective data collection strategies instead of sampling all available events could save data collection costs and correct the bias. Maalouf and Trafalis (2011) implemented kernel logistic regression to rare events data, making use of a fast and robust adaptation of kernel logistic regression and taking the weight of rare events into account. In our method, selective sampling and corresponding prior correction are used to reduce bias induced by rare events.

Since bus GPS records are plentiful and easy to filter, efficient sampling thus can be achieved by creating a balanced dataset in which all bunching events are included and part of the non-bunching events are excluded. A balanced selection to include ones (bunching) and an equal number of zeros (non-bunching) is applied.

Following King and Zeng (2001), prior correction is to correct the estimates according to the fraction of ones in the population, denoted by τ , and the observed fraction of ones in the sample, denoted by $\bar{y}$, since the probability of events to be predicted as ones is overestimated in the sample. The correction is applied to the intercept β_0 as

$$\widehat{\beta}_0 = \beta_0 - \ln \left[\left(\frac{1-\tau}{\tau} \right) \left(\frac{\bar{y}}{1-\bar{y}} \right) \right] \quad (3.8)$$

3.2.5 Linear regression and SVM as benchmark solutions

We now turn to two headway prediction methods that we consider as benchmarks compared to the afore introduced direct bunching prediction method. Firstly, we consider LR (Linear Regression) which is a basic tool in addressing prediction problems. To make LR comparable with LOGR, the same set of independent variables $\mathbf{X}_m^n = [t_m^{n-k}, t_{m-1}^{n-k}, \Delta_{m-1,m}^{n-k}]$ is applied. With $\boldsymbol{\beta}' = [\beta'_0, \beta'_1, \beta'_2, \beta'_3]$ the relationship between the headway at stop n and the set of the independent variables containing information k -stop-ahead is modeled as

$$\Delta_{m-1,m}^n = \boldsymbol{\beta}' \mathbf{X}_m^n \quad (3.9)$$

Secondly, SVM (Support Vector Machine) can map a non-linear relationship for model input and output, and is tested by a number of studies in predicting bus headway or arrival time (Yu et al., 2011; Yu et al., 2016). The same independent variables and dependent variable are applied to the SVM regression, and a RBF (Radial Basis Function) kernel is selected because it is found both efficient for bus arrival time prediction (Yu et al, 2011) and for bus headway prediction (Yu et al, 2016).

3.3 Data processing

Buses are the main mode of public transport in Kyoto, Japan with more than 100 lines being served by several operators. Bus GPS data of two primary bus operators has been obtained for a period of six months in 2016. The data is collected every 8 seconds and provides the geographic coordinates of bus location in real-time as well as associated bus line and vehicle number. Due to the lack of stop-based information, it is essential to identify arrival and departure times for each bus run at each stop. Using bus stop coordinates the distances of a bus from previous and next stops can be computed for every

GPS record. Considering that bunching and traffic congestion might make it difficult for the bus driver to stop the bus at the exact bus stop coordinates as well as inaccuracy of GPS records, the bus is regarded arriving at the stop once it approaches the bus stop within 30m. In the same way the departure time is obtained when the GPS records indicate that the bus has moved 30m from the bus stop.

The data collection period includes the months of April and November. During these months, Kyoto City experiences vast numbers of domestic and foreign visitors who come to enjoy the cherry blossoms (April) and red leaves (November) in various sites around the city. The bus operators thus encounter a huge challenge during these seasons to deliver a reliable service.

A circular bus line, Kyoto City Bus No. 205, which connects the city center, railway station and several famous tourist attractions (Figure 3.1(middle)) is selected for the case study. There are 53 stops on this bus line in total. To exclude the effect of dispatching at the terminal and factors for which we do not have data (e.g. driver issues, departure time adjustments), the 2nd stop of the line is taken as the initial stop and the 52nd stop as the last one so that each bus run passes 51 bus stops. Data of five weekdays in April 2016 are used as the training dataset and those of another five weekdays in the same month are used for testing the model.

The scheduled headway varies from hour to hour, and the mean scheduled headway at the initial stop is 6.97min from 6 am to 8 pm. The shortest scheduled headway is 3min at 7 am. Based on this, 1min is used for the bunching threshold as larger threshold can include headway variance that does not lead to bunching.

Adequate time is required to project a successful correction, in particular, if the control strategy is based on manual communication between the dispatcher and the bus drivers. In this study, the proposed approach is tested under a long prediction horizon of 10 stops or more which gives the operator more than 15min to react since the mean stop-to-stop travel time is 1.77min.

Figure 3.2 illustrates the bus runs departing from the initial stop between 8 am and 10 am. Bunching occurs frequently along the bus line. Bus runs that are involved in bunching as the back bus of two or more buses at least once are denoted in red, and the front buses of a bunching sequence are denoted in blue. Buses in green are not involved in any bunching. The headway fluctuation patterns of seven red

trajectories are demonstrated in Figure 3.3. Because of the bunching effect, the forward-looking headway of back buses fluctuate within a small range, but always below one minute, once bunching has been occurring giving further support to our threshold choice of one minute.

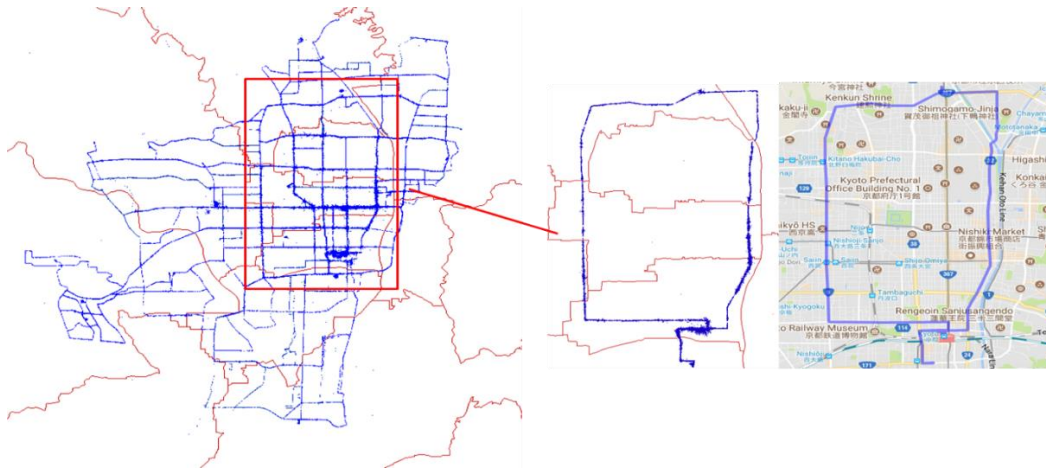


Figure 3.1 Data collected (left), data of Kyoto City Bus No. 205 (middle) and its configuration on real map (right)

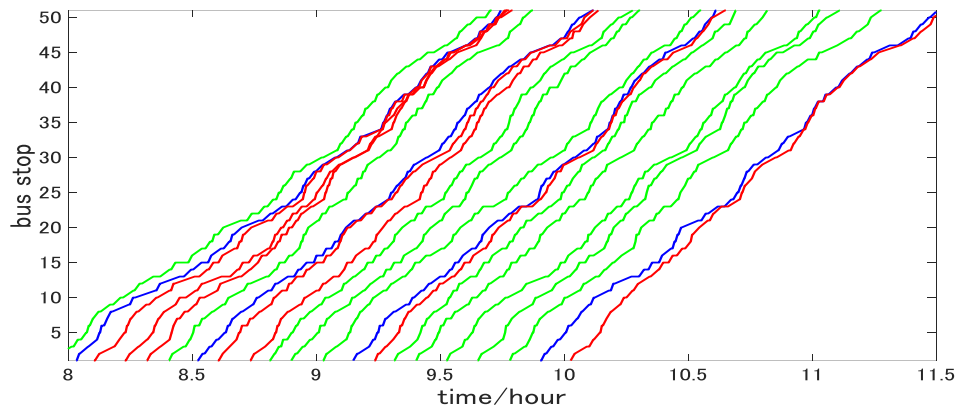


Figure 3.2 Trajectories of Kyoto City Bus No. 205 in one day of April 2016

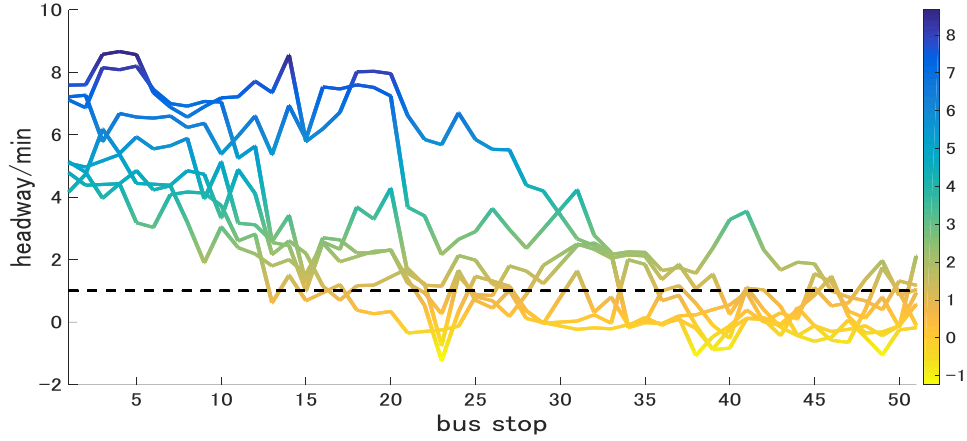


Figure 3.3 Headway fluctuation along the line for bunched buses

3.4 Headway prediction results

In the now following case study, the headway prediction results derived by LR and SVM are discussed at first including a comparison of these results. In Section 3.5 then the focus is on bunching prediction use these two methods as well as the newly proposed LOGR model. In the third part of our case study we compare the measures using ROC curves.

3.4.1 Linear regression results

Table 3.1 shows the estimation results of the fitted LR model. For all the prediction horizons, the headway between the target bus and its front bus $\Delta_{m-1,m}^{n-k}$ and the dwell time of the target bus t_m^{n-k} are always significant at 0.1% level and have positive signs. For short-term prediction the coefficient of $\Delta_{m-1,m}^{n-k}$ is close to 1 and it begins to deviate from 1 as the prediction horizon increases. Meanwhile the coefficient of t_m^{n-k} increases gradually as the prediction horizon extends. t_{m-1}^{n-k} is insignificant in some cases, but it is still considered an important variable indicating at-stop activities and passenger loads. Long t_{m-1}^{n-k} may shorten the headway, but it sometimes results from in-vehicle crowding as well as high boarding demand which may cause boarding failures that lead the following bus to dwell longer and increase the headway, thus the sign of t_{m-1}^{n-k} is inconclusive but mostly negative.

Table 3.1 Coefficients of the independent variables in the LR model

Prediction horizon	Intercept	$\Delta_{m-1,m}^{n-k}$	t_m^{n-k}	t_{m-1}^{n-k}	Adjusted R ²
1-stop-ahead	-0.3604***	1.0009***	0.7084***	0.1247***	0.9681
2	-0.3689***	1.0026***	0.7735***	0.0381*	0.9431
3	-0.4890***	1.0037***	0.9786***	0.0845***	0.9173
4	-0.4664***	1.0051***	0.9611***	0.0203	0.8922
5	-0.5081***	1.0065***	1.0315***	0.0115	0.8673
6	-0.4676***	1.0075***	0.9974***	-0.0751*	0.8424
7	-0.5320***	1.0092***	1.0664***	-0.0247	0.8183
8	-0.5070***	1.0108***	1.0177***	-0.0738*	0.7942
9	-0.5458***	1.0111***	1.1231***	-0.0931*	0.7710
10	-0.5613***	1.0123***	1.1489***	-0.1072**	0.7474
11	-0.5937***	1.0136***	1.1834***	-0.0932*	0.7245
12	-0.6405***	1.0160***	1.2057***	-0.0459	0.7012
13	-0.6283***	1.0186***	1.1944***	-0.0935	0.6790
14	-0.6675***	1.0212***	1.2323***	-0.0886	0.6564
15	-0.6325***	1.0235***	1.2236***	-0.1681***	0.6349

*** ≤ 0.001 , ** ≤ 0.01 , * ≤ 0.05

3.4.2 Support vector machine

The RBF function has two tuning parameters (C , γ) to enhance the predicting power of the SVM model. C is the cost parameter to penalize the misclassifying of a sample. C thus controls the complexity of the classifier; a high C may greatly bend the “prediction hyperplane” to avoid any misclassifying (Cherkassky and Ma, 2004). γ is the inverse of the radius of influence by the samples selected as the support vectors of the model. γ determines the influence of a single sample, a high γ thus may reduce the radius and limit the generalization performance of the model. According to the findings on bus arrival time prediction in Yu et al (2011), $C \in [2^{-5}, 2^5]$, $\gamma \in [0.1, 0.3]$ are recommended for the two parameters. In our method, $(2^2, 1)$ is set for the two parameters after a grid search in which $\gamma = 1$ performs better in our dataset.

3.4.3 Performance evaluation index

MAPE (Mean Absolute Percentage Errors) and RMSE (Root Mean Square Errors) are commonly used to evaluate the prediction performance regarding exact value arrival time or headway prediction. Let M and N denote the total number of bus runs and stops for a bus line, $\Delta_{m-1,m}^{n-k}$ and $\hat{\Delta}_{m-1,m}^{n-k}$ denote the actual value and predicted value for headway, MAPE and RMSE are obtained respectively in Eq. (3.10) and Eq. (3.11). In order to prevent the denominator being close to zero, we follow the method of Yu et al. (2016) to calculate MAPE and use the mean of actual headways $\bar{\Delta}$ instead of $\Delta_{m-1,m}^{n-k}$

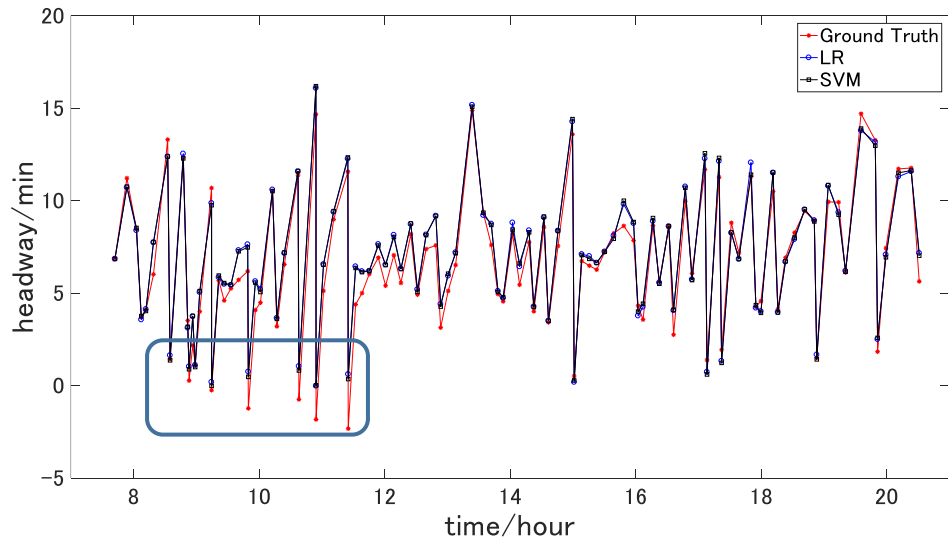
$$MAPE = \frac{1}{(M-1)(N-k)} \sum_{n=k+1}^N \sum_{m=2}^M \left| \frac{\Delta_{m-1,m}^{n-k} - \hat{\Delta}_{m-1,m}^{n-k}}{\bar{\Delta}} \right| \times 100\% \quad (3.10)$$

$$RMSE = \sqrt{\frac{1}{(M-1)(N-k)} \sum_{n=k+1}^N \sum_{m=2}^M (\Delta_{m-1,m}^{n-k} - \hat{\Delta}_{m-1,m}^{n-k})^2} \quad (3.11)$$

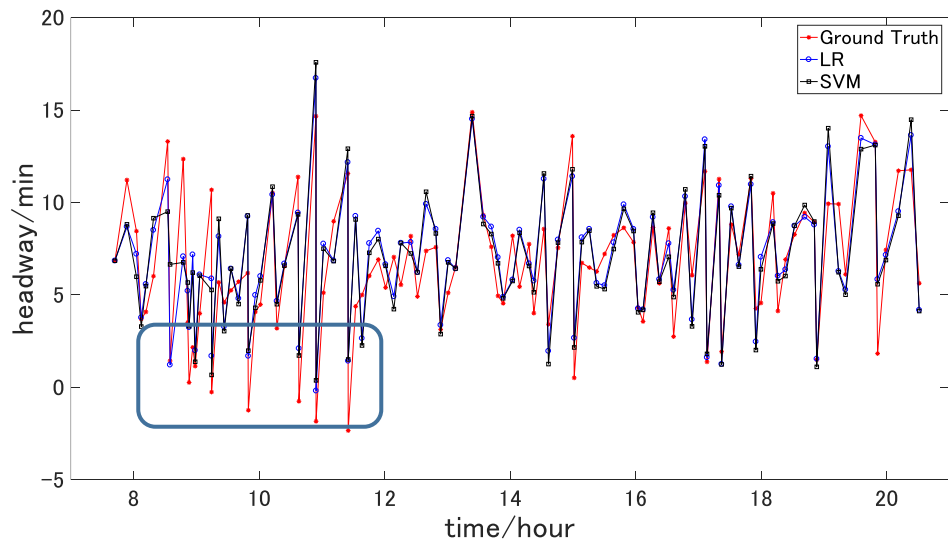
3.4.4 Performance comparison

Headway prediction results at Stop 23 “Kinkaku Temple”, one of the most frequented sightseeing spots in Kyoto, is used to illustrate the performance of the aforementioned two methods. The results of 1-stop-ahead and 10-stop-ahead predictions are illustrated in Figure 3.4 and evaluated in Figure 3.5.

Reliable prediction results (MAPE = 7.42% and RMSE = 0.71min by LR, MAPE = 7.45% and RMSE = 0.71min by SVM) are produced for 1-stop-ahead prediction. For 10-stop-ahead prediction, the results obviously deteriorate (MAPE = 21.64% and RMSE = 1.93min by LR, MAPE = 21.51% and RMSE = 1.92min by SVM). We suggest they can still provide insights into expected fluctuation patterns downstream, but the exact value is not reliable. Furthermore, neither in 1- nor 10-stop-ahead prediction can these two methods perform favorably under the circumstance that the actual headway becomes extremely short and bunching is going to happen, as is highlighted by the blue box in Figure 3.4. Furthermore, Figure 5 illustrate that in terms of MAPE and RMSE, both methods produce close prediction accuracy and deteriorate similarly. Instead of significant increases in prediction errors, evaluation metrics deteriorate gradually as the prediction horizon extends.

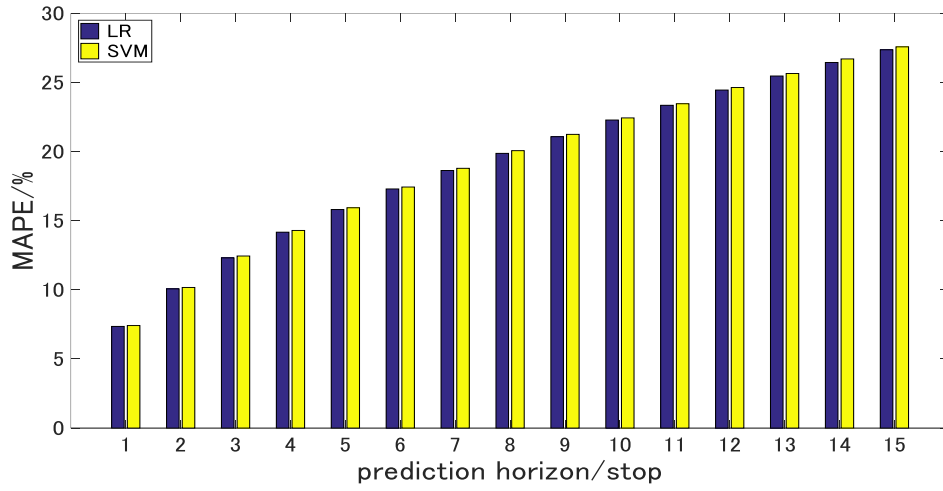


(a) 1-stop-ahead headway prediction

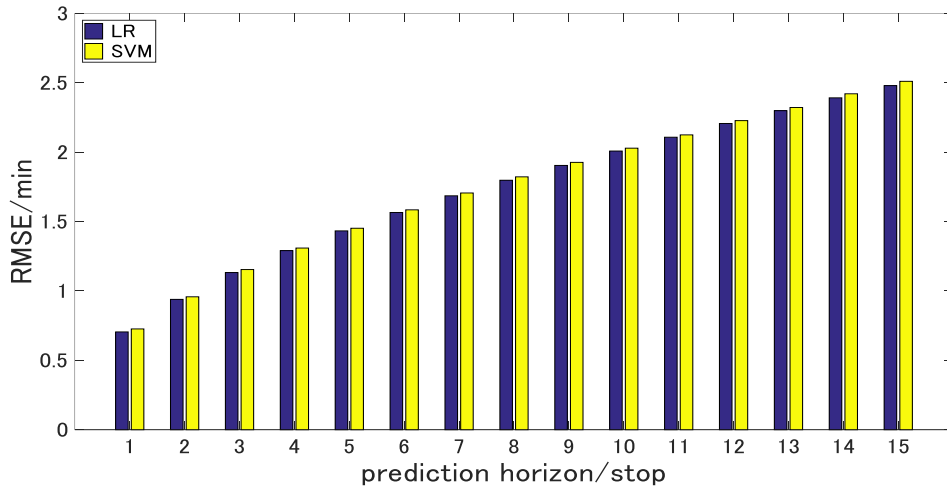


(b) 10-stop-ahead headway prediction

Figure 3.4 Performance comparison in terms of exact headway value



(a) Deterioration in MAPE as the prediction horizon extends



(b) Deterioration in RMSE as the prediction horizon extends

Figure 3.5 Performance comparison in terms of RMSE and MAPE under various prediction horizons

3.5 Bunching prediction

3.5.1 Logistic regression results

We now focus on bunching prediction, firstly with logistic regression. Estimation results with and without rare events bias correction are shown in Table 3.2. Adjusted McFadden's R^2 obtained by Eq. (3.12) is selected to measure the overall goodness of fit for the logistic regression model.

$$R_{MCF}^2 = 1 - \frac{\ln L_{full} - K}{\ln L_{null}} \quad (3.12)$$

where L_{full} is the likelihood derived by the fitted model, and L_{null} is the likelihood of a null model with intercept as the only predictor. K is the number of independent variables in the proposed model. Due to the randomness generated by drawing non-bunching observations from the 5-day dataset to correct the rare event bias, we run the model for 100 times and report the mean values for the coefficients and adjusted McFadden's R^2 . The significance is not based on any specific run but on all the 100 runs, and for each variable the p-value is obtained by one sample test on the 100 estimated coefficients. Bunching probability is negatively correlated with the value of headway, thus the coefficients of the variables in the fitted LOGR have a reversed sign compared to those in the LR model. The correction is proven effective as the adjusted R_{MCF}^2 is increased by at least 0.05 for each prediction horizon.

Table 3.2 Coefficients of the independent variables in the LOGR model

Prediction horizon	Intercept	$\Delta_{m-1,m}^{n-k}$	t_m^{n-k}	t_{m-1}^{n-k}	Adjusted R_{MCF}^2
Without correction					
1-stop-ahead	3.0842***	-2.2341***	-1.4015***	-0.3552***	0.7508
2	2.3653***	-1.7302***	-1.0123***	-0.1913*	0.6899
3	2.0826***	-1.4265***	-1.1577***	-0.1057	0.6357
4	1.8370***	-1.2634***	-1.0737***	0.0149	0.6016
5	1.6296***	-1.1234***	-0.9013***	-0.0036	0.5651
6	1.5044***	-1.0363***	-0.9266***	0.1282	0.5385
7	1.4301***	-0.9596***	-0.8913***	0.1007	0.5097
8	1.2780***	-0.8957***	-0.7406***	0.1680*	0.4838
9	1.2385***	-0.8352***	-0.8607***	0.1960**	0.4558
10	1.1162***	-0.7811***	-0.8076***	0.2664***	0.4289
11	1.1189***	-0.7361***	-0.7486***	0.0807	0.4028
12	1.0504***	-0.7003***	-0.7425***	0.1446*	0.3819
13	0.9763***	-0.6679***	-0.6612***	0.1593*	0.3615
14	0.9184***	-0.6403***	-0.5862***	0.1666**	0.3431
15	0.8498***	-0.6137***	-0.5155***	0.1807**	0.3246
With correction (mean values of 100 runs reported with significance also based on all runs)					
1-stop-ahead	2.3647***	-2.0343***	-1.1165***	0.0619***	0.8214
2	1.8730***	-1.5919***	-0.8407***	0.0385*	0.7732
3	1.7571***	-1.3429***	-1.1465***	0.0985***	0.7267
4	1.6035***	-1.2226***	-0.9988***	0.1627***	0.6971
5	1.4231***	-1.0874***	-0.9060***	0.1598***	0.6568
6	1.2997***	-1.0042***	-0.8988***	0.2585***	0.6248
7	1.2399***	-0.9326***	-0.7756***	0.1574***	0.5938
8	1.1046***	-0.8738***	-0.7128***	0.2740***	0.5641
9	1.1111***	-0.8146***	-0.8855***	0.2514***	0.5311
10	1.0183***	-0.7663***	-0.8584***	0.3289***	0.5023
11	0.9848***	-0.7188***	-0.7544***	0.1642***	0.4709
12	0.9108***	-0.6817***	-0.7792***	0.2616***	0.4461

Prediction horizon	Intercept	$\Delta_{m-1,m}^{n-k}$	t_m^{n-k}	t_{m-1}^{n-k}	Adjusted R_{MCF}^2
13	0.8268***	-0.6467***	-0.7202***	0.2776***	0.4195
14	0.8020***	-0.6183***	-0.6655***	0.2246***	0.3970
15	0.6966***	-0.5887***	-0.5887***	0.2788***	0.3739

*** ≤ 0.001 , ** ≤ 0.01 , * ≤ 0.05

3.5.2 Performance evaluation index

We define an actual bunching as “observed positive” and a predicted bunching as “predicted positive”. Similarly, for non-bunching we define “observed negative” and “predicted negative”. All the prediction results can be cast into four categories as is shown in Table 3.3, e.g. it is a true positive if an observed bunching is correctly labeled one in the prediction outcomes. Four indices can be obtained from Eq. (3.13) to Eq. (3.16). A binary classifier with high true positive rate and high true negative rate is desired. The former is commonly referred to as “sensitivity” and the latter as “specificity”. Sensitivity, specificity and accuracy, which is an index computed with Eq. (17) to indicate overall prediction performance, are applied to evaluate the binary classification performance of the three algorithms.

Table 3.3 Four categories for binary classification results

	Observed positive (OP)	Observed negative (ON)
Predicted positive (PP)	True positive (TP)	False positive (FP)
Predicted negative (PN)	False negative (FN)	True negative (TN)

$$\text{True positive rate (TPR, sensitivity, SES)} = \frac{\sum TP}{\sum OP} \quad (3.13)$$

$$\text{False positive rate (FPR)} = \frac{\sum FP}{\sum ON} \quad (3.14)$$

$$\text{True negative rate (TNR, specificity, SPC)} = \frac{\sum TN}{\sum ON} \quad (3.15)$$

$$\text{False negative rate (FNR)} = \frac{\sum FN}{\sum OP} \quad (3.16)$$

$$\text{Accuracy (ACC)} = \frac{\sum TP + \sum TN}{\sum OP + \sum ON} \quad (3.17)$$

For headway-based methods, only one combination of sensitivity and specificity is derived, as headway prediction produces an exact value for each headway, resulting in deterministic true positive and negative outcomes. Instead, by using logistic regression different combinations are obtained depending on the cut-off point applied to the predicted probability. The cut-off point is the threshold to determine the predicted positive. The event is judged as positive if its predicted probability exceeds the cut-off point. A high cut-off point tends to only identify events presenting convincingly high probability as positives, and consequently, it thus might misclassify observed positives as negative. Vice versa, a low cut-off point will lead to more false positives. Therefore the cut-off point choice should depend on the operator's attitude towards bunching. Two scenarios are assumed here to represent operators with different weights to false negative errors (missing actual bunching). Moreira-Matias et al (2016) employed a large weight of 10:1 for false negative compared to false positive for aggressive control purposes. We consider more moderate weights of 1:1 and 3:1.

Scenario 1 (LOGR-N): the operator is bunching-neutral, and gives equal weight to false positive and false negative.

Scenario 2 (LOGR-A): the operator is bunching-averse, and gives a 3:1 weight to false negative over false positive predictions.

The cost function in Eq. (3.18) computes the total weighted errors given a cut-off point. For LOGR-N, $w_{FP} = w_{FN} = 1$, and for LOGR-A, $w_{FP} = 1$, $w_{FN} = 3$. The cut-off point generating the lowest cost is taken as the optimal one. Based on the scenario-specific predicted positives and negatives, the combination of sensitivity and specificity is determined.

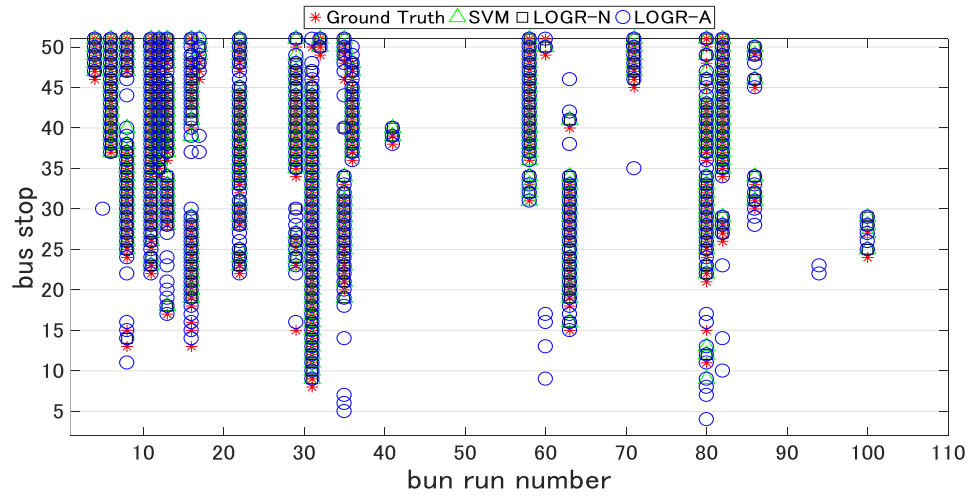
$$c = w_{FP} \sum FP + w_{FN} \sum FN \quad (3.18)$$

3.5.3 Performance comparison

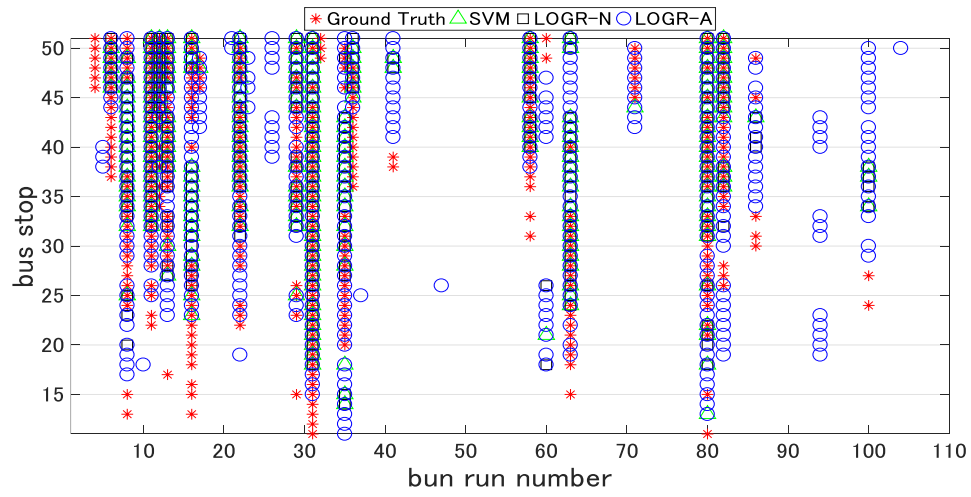
Considering that the results derived by LR and SVM are similar, the comparison here is among SVM and two distinguished scenarios based on LOGR. As is presented in Figure 3.6(a), most bunching events can be detected 1-stop in advance by all three methods, and LOGR-A produces several false

positives because it applies a more aggressive strategy to potential bunching events. However, LOGR-A significantly outperforms in 10-stop-ahead prediction, as is illustrated in Figure 3.6(b). LOGR-A captures a number of observed positives that are misclassified by SVM and LOGR-N although it generates a few more false positives.

A further comparison among two headway-based approaches and two scenarios of logistic regression is demonstrated in Figure 3.7. Sensitivity, specificity and accuracy for the four methods under various prediction horizons are presented. LOGR-A shows remarkable robustness in terms of sensitivity. On the contrary to the obvious deterioration of the other three methods, the sensitivity of LOGR-A keeps above 65% under all the prediction horizons. Besides, it only slightly underperforms the other three methods in terms of specificity, indicating an acceptable trade-off cost. Non-bunching events overwhelm bunching events in the daily operation, and a slight underperformance in specificity might introduce a large number of false positive. The exact numbers of true positives, false positives, true negatives, false negatives derived in the 5-day testing data are listed in Table 3.4. LOGR-N always generates the least total errors (highest accuracy). LOGR-A always correctly detects most bunching events (highest sensitivity) at the cost of most total errors (lowest accuracy). The notable advantage of LOGR over the other two methods is its trade-off functionality. It can achieve highest overall accuracy and can outperform the other methods in terms of sensitivity, although it cannot realize both objectives simultaneously.

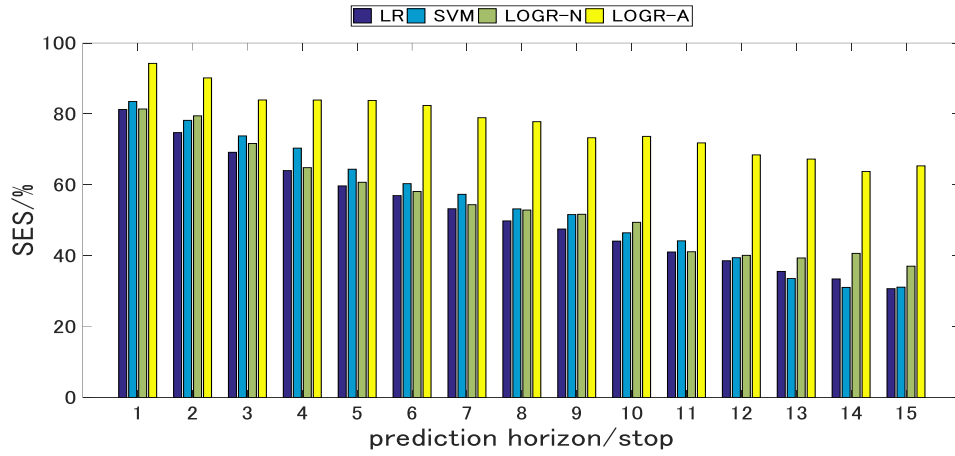


(a) 1-stop-ahead bunching prediction

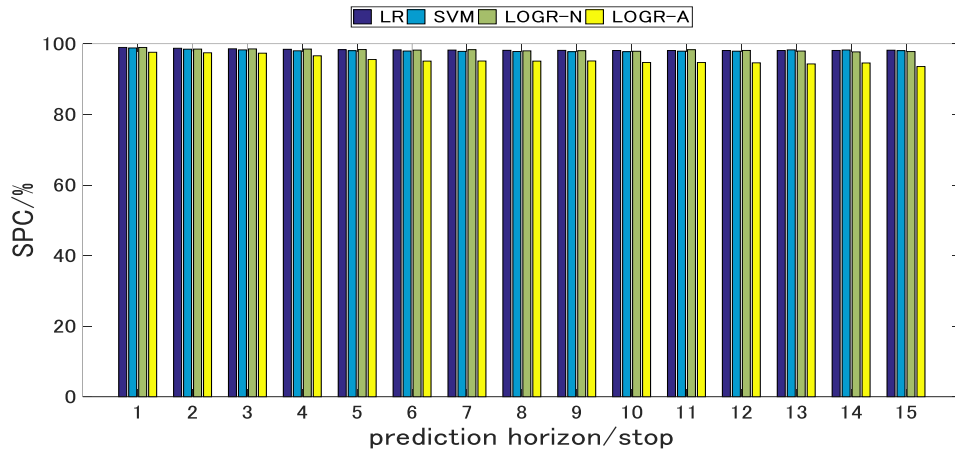


(b) 10-stop-ahead bunching prediction

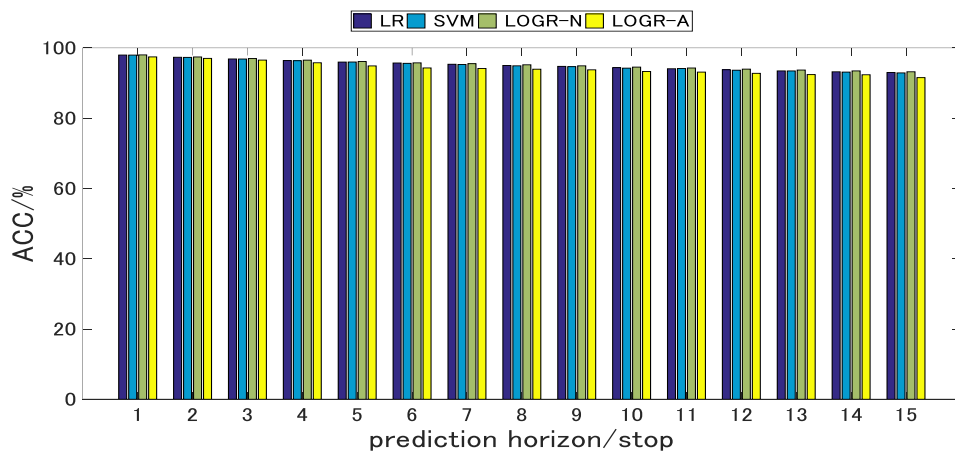
Figure 3.6 Performance comparison in terms of binary bunching identification



(a) Deterioration in SES as the prediction horizon extends



(b) Deterioration in SPC as the prediction horizon extends



(c) Deterioration in ACC as the prediction horizon extends

Figure 3.7 Performance comparison in terms of SES, SPC and ACC under various prediction horizons

Table 3.4 Performance comparison for 10-stop-ahead bunching prediction

	Size	OP	PP		PN		SES	SPC	ACC
			TP	FP	TN	FN	(%)	(%)	(%)
Day 1									
LR	4223	344	161	77	3802	183	46.80	98.01	93.84
SVM	4223	344	171	89	3790	173	49.71	97.71	93.80
LOGR-N	4223	344	208	118	3761	136	60.47	96.96	93.99
LOGR-A	4223	344	262	200	3679	82	76.16	94.84	93.32
Day 2									
LR	4182	384	158	118	3680	226	41.15	96.89	91.77
SVM	4182	384	166	138	3660	218	43.23	96.37	91.49
LOGR-N	4182	384	143	94	3704	241	37.24	97.53	91.99
LOGR-A	4182	384	266	345	3453	118	69.27	90.92	88.93
Day 3									
LR	4264	355	175	62	3847	180	49.3	98.41	94.32
SVM	4264	355	187	73	3836	168	52.68	98.13	94.35
LOGR-N	4264	355	201	82	3827	154	56.62	97.9	94.47
LOGR-A	4264	355	263	202	3707	92	74.08	94.83	93.11
Day 4									
LR	4182	146	51	42	3994	95	34.93	98.96	96.72
SVM	4182	146	57	51	3985	89	39.04	98.74	96.65
LOGR-N	4182	146	50	37	3999	96	34.25	99.08	96.82
LOGR-A	4182	146	100	112	3924	46	68.49	97.22	96.22
Day 5									
LR	4182	254	123	67	3861	131	48.43	98.29	95.27
SVM	4182	254	121	79	3849	133	47.64	97.99	94.93
LOGR-N	4182	254	176	115	3813	78	69.29	97.07	95.38
LOGR-A	4182	254	204	164	3764	50	80.31	95.82	94.88

3.6 Discussion on the trade-off between sensitivity and specificity

ROC curves created by plotting (1-SPC, SES) for given cut-off points are commonly used to evaluate the classification performance. AUC (Area Under the Curve) being close to one indicates good classification power. ROC curves under various prediction horizons are presented in Fig. 3.8. Furthermore, the four combinations of sensitivity and specificity derived by the four methods discussed in the previous section are indicated on each curve.

For each horizon, the corresponding curve can be considered the optimal front derived by LOGR. If an algorithm outperforms LOGR, the point it represents should appear above the curve with a higher SES and lower 1-SPC. It can be observed that the two headway-based methods (LR and SVM) generally fall below and sometimes on the LOGR curve, although the downward deviation from the curve is not significant.

It is easy to conduct the trade-off between sensitivity and specificity on the LOGR curve. The LOGR curve contains all combinations of prediction performance given continuous cut-off points where each cut-off point can be considered as optimal. A bunching-averse operator who is aggressive to eliminate bunching might desire to detect 99% of the positives regardless of the cost to increase false positive rate. This trade-off functionality significantly enhances the flexibility and robustness of existing bunching prediction approaches, especially for putting the predictive methodology into real practice. The curves provide a robust benchmark and insights for future algorithms that address bunching prediction problem. Deterministic methods can only produce one combination of prediction performance which greatly limits its contribution to the real application unless its sensitivity and specificity simultaneously achieve a highly reliable level. Other probabilistic methods generating a curve having higher AUC than LOGR or deterministic methods producing points of substantial upward deviation from the curve under various prediction horizons should be further promising extensions.

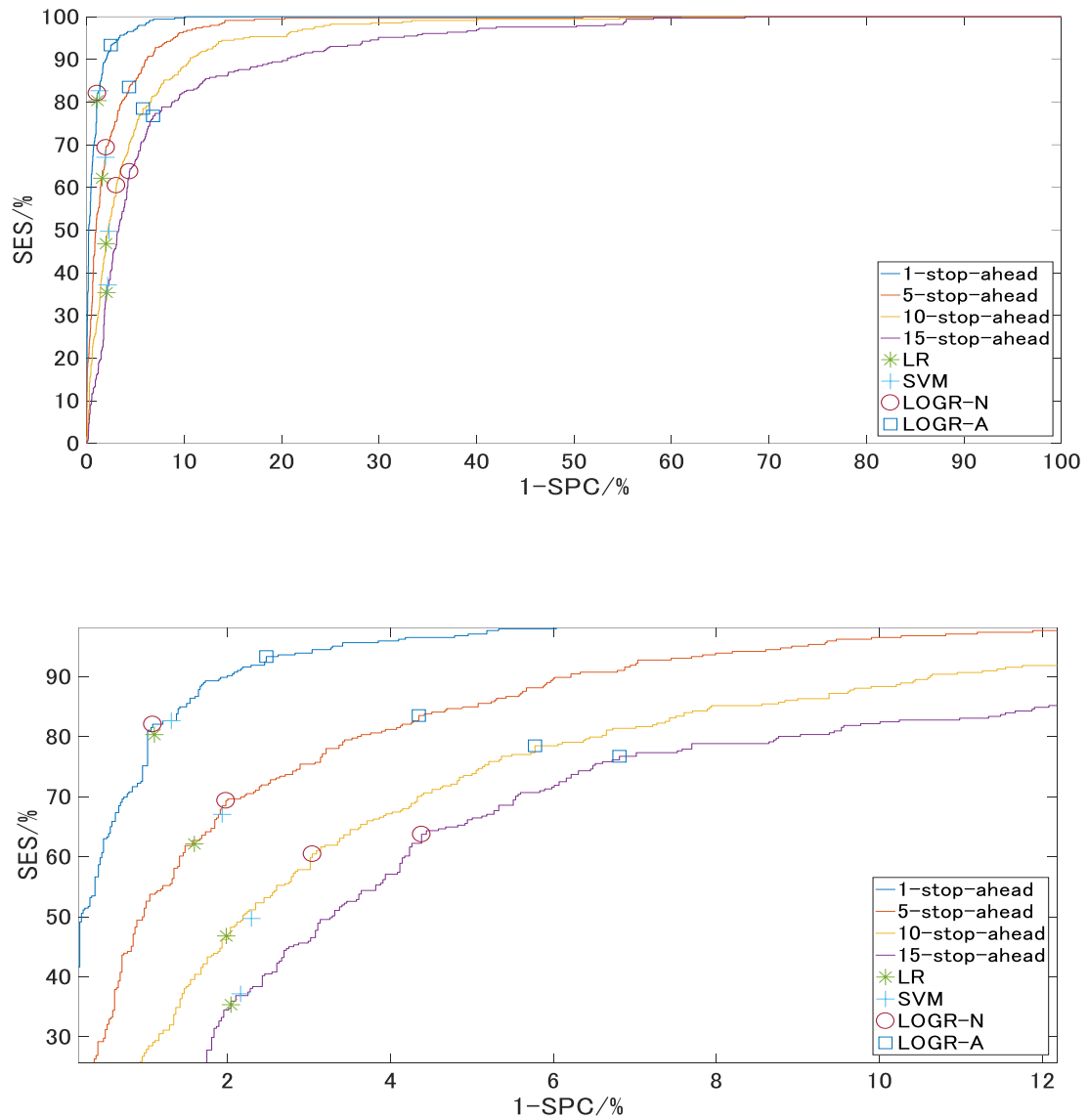


Figure 3.8 ROC curves under various prediction horizons (1-stop, 5-stop, 10-stop and 15-stop ahead)

Table 3.5 Supplementary information for ROC curves shown in Figure 3.8

Prediction horizon	AUC (area under the curve)	Cut-off point of LOGR-N (%)	Cut-off point of LOGR-A (%)
1-stop-ahead	0.9922	90.31	73.5
5	0.9763	87.19	77.53
10	0.9546	87.56	79.74
15	0.9279	81.6	71.50

3.7 Summary

In this chapter, the potential of logistic regression to predict bus bunching is discussed. We consider the “rare event” nature of our problem which leads logistic regression to lose prediction power due to being biased to the majority in the dataset where positive events are by far outnumbered by negative events. Thus a selective sampling method and intercept correction is applied. We then compare this method with existing approaches that predict headways and then utilize the headway prediction for bunching prediction. Clearly headway prediction can be used for a larger range of purposes and deeper understanding of the service regularity developments as well as control strategies. However, bunching prediction itself is important as it can be considered a distinctive state. This research and other literature illustrate that headways fluctuate, but that, once bunching is reached, this state mostly continues along the line with far less headway fluctuation. We illustrate that when it comes to predicting bunching itself the newly proposed method has the potential to outperform headway-based methods such as LR and SVM in several aspects.

Firstly, LOGR provides superior prediction results under a long prediction horizon. It outperforms LR and SVM by 28% in sensitivity and maintains the same level of specificity in 10-stop-ahead prediction. It also shows improved resistance against deterioration in prediction performance as the prediction horizon extends.

Secondly, robustness and flexibility are significantly enhanced. LOGR provides robust prediction results that contain various sets of bunching outcomes under different cut-off points. This enables the operator to apply weights that are in accordance with their attitude towards bunching and operation budget. Some operators with limited possibility or willingness to apply corrective measures can use SVM or LOGR with neutral cut-off point setting. On the contrary, operators who desire to eliminate any possible bunching might be unwilling to choose headway-based methods which omit a considerable number of bunching in the long-term prediction cases. In this case LOGR-A becomes a much-preferred option. To conclude, LOGR provides operators with a wide range of options that can be tailored by their attitudes towards unexpected system disturbances.

We find that the headway-predicting approaches deviate slightly downward from the optimal front and we discuss that their shortcomings are inadequate robustness and flexibility from the operator’s

perspective. We note that it is also feasible to form a curve in terms of sensitivity and specificity for probabilistic headway prediction methods with confidence intervals to realize the trade-off on the curve discussed in this research. Hans et al (2015) developed a simulation-based prediction tool, Yu et al (2017) tested RVM algorithm on headway prediction problem. Both methods could be extended to compute the probability of a headway falling below 1min and then to construct the ROC curves. By doing so and comparing the different ROC curves more insights might be obtained.

Finally, other extensions that potentially strengthen the predictive power of the models presented in this study should be noted. The model itself has space for improvement by including variables such as weather, traffic signals and passenger demand that are not incorporated due to missing data. Furthermore, the study could be extended to simultaneously predict bunching for several lines, in which case common line effects such as the interaction between buses of different lines at a common stop need to be considered.

Chapter 4 **Demand estimation using bus AVL data**

4.1 Introduction

In this chapter, we explore the feasibility of estimating the passenger origin-destination (OD) flows for a bus transit route using Automatic Vehicle Location (AVL) data. Dwell time models are used to build connections between bus dwell times and passenger OD flows, and a modified gravity model is applied to reduce the number of unknown parameters. Bayesian inference and Markov Chain Monte Carlo (MCMC) methods are implemented to estimate the mean and confidence intervals for the boarding-alighting flows and passenger loads at each stop. The methodology is validated by Automatic Fare Collection (AFC) data. It is found in the case study that the estimation performance derived by using bus AVL data matches that using unchained passenger boarding-alighting counts. Furthermore, additional input data that roughly characterize the importance of bus stops improve the performance of OD estimation. We conclude that our proposed methodology can support bus transit operators having no or limited information regarding passenger demand in planning and operation aspects.

Due to rapid advancements in data collection possibilities, nowadays some transit operators are in a position to conduct data-driven route configuration planning, scheduling and capacity bottleneck estimation, as well as to execute real-time controls. Fluctuations in passenger demand can be accommodated, which then substantially improve the service quality and enhance the competitiveness of traditional bus services against other emerging transport modes. However, it should not be overlooked that a large number of operators have still limited access to data sources such as smart card data.

Here we consider the problems from the perspective of such a transit operator who has no direct observation on the passenger flows. As a result, it is difficult for the operator to plan services satisfying real demand patterns well leading to, for example, unreasonable bus schedules and route and in-vehicle overcrowding. It should be noticed that this dilemma results not only from budget inadequacies or technical difficulties in terms of introducing Automatic Fare Collection (AFC) and Automatic Passenger Count (APC) systems to the city but also from the inefficiencies in data sharing or conflicts of interest, e.g. a unified smart card is issued by a different operator or a third company which does not allow the data to be shared with the bus operators.

In order to help this kind of “data-poor” operator, we propose that dwell times are indirect but informative observations regarding passenger flows, which can be collected by Automatic Vehicle Location (AVL) system. For each bus only one - fairly low cost - GPS device is required to collect the data. As a large set of OD patterns are feasible for each set of dwell times, the problem is highly underspecified and we implemented Bayesian inference approaches which are proven efficient to obtain confidence intervals given uncertain observations (Maher, 1983; Hazelton, 2001; Hazelton, 2008; Hazelton, 2010).

This study therefore significantly extends the role that bus AVL data can play in improving the quality of bus services. AVL systems are originally designed to monitor the punctuality performance by recording the vehicle’s temporal-spatial coordinates. As an extension, bus AVL datasets are further widely used to predict bus arrival times and headways (e.g. Yu, 2011; Hans, 2015). This study illustrates the potential of bus AVL data in estimating passenger flows. The main contribution of this study is hence a novel methodology to estimate two levels of information in terms of passenger demand using bus AVL data as the main data source. It estimates passenger boarding/alighting numbers and onboard numbers for each bus stop of a bus route. Further, at a more disaggregated level, it estimates the stop-based passenger OD flows.

This reminder of this chapter is as follows. Section 4.2 proposes the model framework. Section 4.3 introduces the basic dwell time models incorporated in this model. Section 4.4 builds a Bayesian inference model. The MCMC sampling algorithm is developed in Section 4.5. The case study is described in Sections 4.6 and 4.7. We test the proposed estimation methodology and algorithm on two main bus routes in Shizuoka City, Japan. In Section 4.6, we introduce the two datasets used in the case study: AVL data and AFC data, and the processing methods. In Section 4.7, we evaluate the estimation performance by comparing the results with AFC data. We also investigate the effect of additional observations to AVL data e.g. passenger count data, and some prior knowledge about the estimation performance by distinguishing 6 scenarios. Finally, we discuss the findings of this study and limitations in Section 4.8 and 4.9 respectively.

4.2 Model framework

The notation used in this section is as follows grouped by observations and directness of estimation.

Observations (data)

$D_{i,k}$	dwel time of bus run k at stop i
$W_{i,k}$	passenger activity time of bus run k at stop i
$\Delta_{i,k}$	headway between bus run k and its front bus at the arrival of stop i
t^a	average alighting time per passenger
t^b	average boarding time per passenger
c	vehicle activity time

Estimated Parameters

p_i^g	generation power of bus stop i
p_i^a	attraction power of bus stop i
α	cost function parameter
σ_W	standard deviation of passenger activity time

Directly derived parameters

$a_{i,j}$	arrival rate per minute of passenger boarding at stop i and alighting at stop j
-----------	---

“2nd level” derived parameters

$Q_{i,j,k}$	number of passenger boarding bus run k at stop i and alighting at stop j
$O_{i,k}$	number of passenger on-board when bus run k arrives at stop i
$B_{i,k}$	number of boarding passenger of bus run k at stop i
$A_{i,k}$	number of alighting passenger of bus run k at stop i
$\mu_{i,k}$	expected passenger activity time of bus run k at stop i

For comparison purposes in Section 6, we also estimate the OD flows using models with more information, i.e. if **A** and/or **B** available. In these cases, **A** and/or **B** are moved to the observations.

For each bus run $k = 1, 2, \dots, K$ at each bus stop $i = 1, 2, \dots, N$, we let $D_{i,k}$ denote the bus dwell time, and let a constant c simplify the vehicle activity time which is the aggregation of time required for door opening, door closing, deceleration and acceleration but excluding “passenger activity time”, then the passenger activity time can be written as

$$W_{i,k} = D_{i,k} - c \quad (4.1)$$

Let t^a and t^b denote the average alighting and boarding time per passenger, let $B_{i,k}$ and $A_{i,k}$ denote the boarding flows and alighting flows at stop i for bus run k respectively. With this we can obtain the expected passenger activity time as a function of boarding and alighting flows as well as time required per boarding and alighting process as in Eq. (4.2). The specification of this function depends on the payment process and vehicle layout as discussed in Section 3.2.

$$\mu_{i,k} = f(t^b B_{i,k}, t^a A_{i,k}) \quad (4.2)$$

The passenger activity time observed in real world will fluctuate due to passengers' heterogeneity (e.g. age, payment method, being disabled or not), the friction effect among passengers as well as other passenger and vehicle related accidental events (Lin and Wilson, 1992; Tirachini, 2013; Sun et al, 2014). In order to capture the variation in passenger activity time but to avoid complicating the dwell time model in Eq. (4.2), we assume the passenger activity time to be characterized by a normal distribution as in Eq. (4.3)

$$W_{i,k} \sim N(\mu_{i,k}, \sigma_W^2) \quad (4.3)$$

where σ_W is the standard deviation of passenger activity time. The stochastic activity times are observable in that they are the products of the dwell times which are directly observed from bus AVL data excluding the vehicle activity time which can be inferred according to bus driver's experience or a simple in-vehicle survey. Therefore, the passenger activity times are taken as the random variables based on which Bayesian inference is applied.

For each bus run k , let arrival rate $a_{i,j}$ ($i, j = 1, 2, \dots, N; i < j$) denote the arrival number per minute of passengers who board at stop i and alight at stop j . We assume that the actual arrival number follows a Poisson distribution so that the expected number of passenger that board at stop i and alight at stop j accumulated over the headway $\Delta_{i,k}$ can be obtained as Eq. (4.4).

$$Q_{i,j,k} \sim \text{Poisson}(a_{i,j} \Delta_{i,k}) \quad (4.4)$$

in which we assume that $\Delta_{i,k} > 0$ always holds and newly arriving passengers during the dwell time do not board the bus. The former assumption is realistic if overtaking between the buses at stops is not allowed, which is the case in many cities including our case study city, and a small though close-to-zero time interval exists between the arrival times of two bunched buses. The latter assumption is unnecessary if departure-to-departure headway is employed. We use arrival-to-arrival headway as the bus arrival times in our AVL data are more accurate than the departure ones

As is mentioned in Section 2, the focus of this chapter is to estimate mean (expected) OD flows for a bus route in an observational period instead of reconstructing the OD matrix for each bus run. We thus simplify the number of passengers travelling from i to j in Eq. (4.5), by assuming that the fluctuation in the OD matrix of individual bus runs in a certain time period (morning peak, evening peak, off-peak hours) is all attributed to the endogenous (due to timetable) or exogenous (due to stochastic reasons such as random delay) variations in headway. $a_{i,j}$ are the unknown parameters and $\Delta_{i,k}$ are obtained from bus AVL data.

$$Q_{i,j,k} = a_{i,j} \Delta_{i,k} \quad (4.5)$$

The aggregated boarding flows at stop i and alighting flows at stop j can be derived with Eqs. (4.6) and (4.7) accordingly.

$$B_{i,k} = \sum_{j=i+1}^N Q_{i,j,k} \quad i = 1, 2, \dots, N-1 \quad (4.6)$$

$$A_{i,k} = \sum_{j=1}^{i-1} Q_{j,i,k} \quad i = 2, 3, \dots, N \quad (4.7)$$

The number of passengers on-board when bus run k arrives at stop i can be obtained as Eq. (4.8). Note that $B_{N,k}$, $A_{1,k}$, $O_{1,k}$ are zeros in that no passenger would board at the last stop or alight at the first stop, and there is no passenger when the bus arrives at the first stop.

$$O_{i,k} = O_{i-1,k} + B_{i-1,k} - A_{i-1,k} \quad i = 2, 3, \dots, N \quad (4.8)$$

A modified gravity model is implemented to decompose the arrival rate matrix and thus reduce the number of unknown parameters to estimate. Let p_i^g ($i = 1, 2, \dots, N - 1$) and p_i^a ($i = 2, 3, \dots, N$) denote the generation and attraction power for each bus stop i , so that the arrival rate $a_{i,j}$ can be written as Eq. (4.9). Noting that we have constraints $p_N^g = p_1^a = 0$, we exclude them from the generation and attraction power vectors. In the denominator, the first term denotes the general distance deterrence. The second term is added to acknowledge that passengers seldom travel only one or two stops with α as a parameter to be calibrated.

$$a_{i,j} = \frac{p_i^g p_j^a}{(j-i) + \frac{\alpha}{j-i}} \quad i = 1, 2, \dots, N-1; j = 2, 3, \dots, N; i < j \quad (4.9)$$

Accordingly, the two vectors $\mathbf{p}^g$ and $\mathbf{p}^a$, each of size $N - 1$, and parameter α substitute the arrival rate $a_{i,j}$ as the new unknown parameters. The number of unknown parameters to estimate is hence reduced from $N(N - 1)/2$ to $2(N - 1) + 1$. Together with σ_w , we have $2N$ unknown parameters to estimate in total.

Dwell time $D_{i,k}$ and headway $\Delta_{i,k}$ are directly observed from bus AVL data. Average boarding/alighting time per passenger and vehicle activity time can be obtained by an onboard survey in a small sample or according to the bus driver's experience. Arrival rate $a_{i,j}$ are the derived parameters dependent on the estimated parameters. OD flows $Q_{i,j,k}$, boarding flows $B_{i,k}$, alighting flows $A_{i,k}$ and passenger loads $O_{i,k}$ are thus the “2nd level” derived parameters dependent on the estimated arrival rate $a_{i,j}$ and observed headway $\Delta_{i,k}$. The relationships between the unknown parameters and the observations are illustrated in Figure 4.1. We would like to point out that the OD flows are the main estimation objective though, boarding/alighting flows and passenger loads are also estimated by the proposed framework, which can be regarded as the lower-level objective and substantially benefit the “data-poor” operators with inferred stop-based passenger dynamics and capacity bottlenecks.

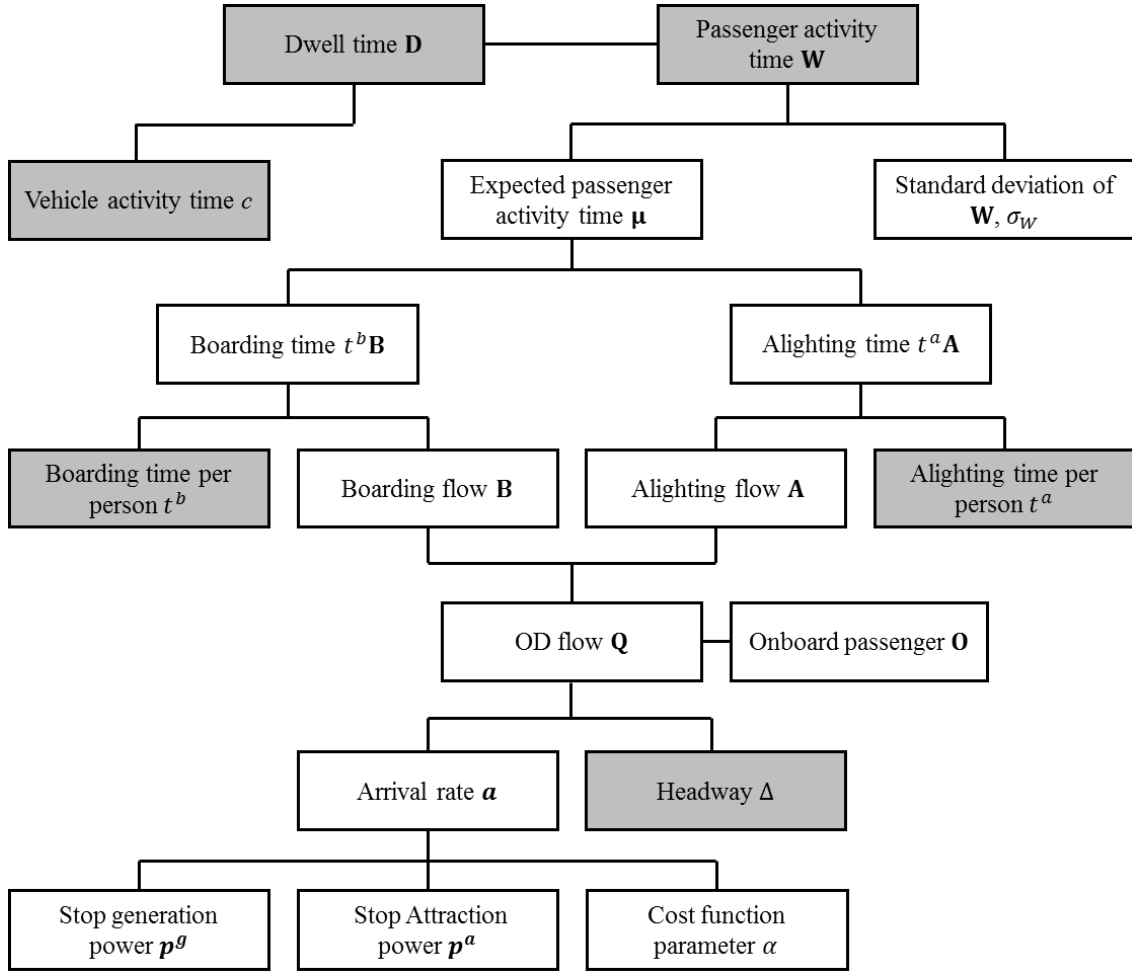


Figure 4.1 Model framework; Items in grey boxes are the observations and data input; Items in white boxes are the unknown parameters and derived parameters

4.3 Basic dwell time models

We now turn to the specification of the dwell time function in Eq. (4.2). Generally sequential and simultaneous boarding and alighting can be distinguished and lead to different formulations of the expected passenger activity time as in Eq. (4.10).

$$\mu_{i,k} = \begin{cases} \max(t^b B_{i,k}, t^a A_{i,k}) & \text{simultaneous boarding and alighting} \\ t^a A_{i,k} + t^b B_{i,k} & \text{sequential boarding and alighting} \end{cases} \quad (4.10)$$

The choice of dwell time models is flexible in our proposed estimation framework. We select the simultaneous form as it fits the situation in the case study city. We note that in general the sequential form is, however, easier to estimate as the dwell time becomes less ambiguous and thus the solution

space is smaller. For the transit system where boarding always starts after alighting, the sequential form is supposed to be more suitable. Lin and Wilson (1992) applied the sequential form on a single door, and the simultaneous form on the multiple doors of a train vehicle. Other more complex dwell time models distinguish the average boarding/alighting time for the passenger of different personal attributes. Tirachini (2013) considered the effect of different methods of payment and age on the average boarding/alighting time. Sun et al. (2014) reviewed and tested a range of dwell time specifications by building both linear and nonlinear regression models to explain the average boarding/alighting time with consideration of the in-vehicle overcrowding, and found that average boarding/alighting time tends to be negatively correlated to the number of total boarding/alighting passengers at a stop. They also found that the friction effect significantly delays the boarding activity and turns a simultaneous case into a sequential case when 63% of the total capacity is occupied by the passengers. These advanced dwell time models are not incorporated in this chapter, as we try to avoid introducing more unknown parameters or increasing the structural complexity of the dwell time model for this underspecified estimation problem.

4.4 Bayesian inference

In this section, we apply Bayes' theorem to infer the posterior probability distribution for the unknown parameters. We remind that the observations used are the passenger activity times $\mathbf{W}$ and that the set of unknown parameters is $\{\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_w\}$.

Let $f(\mathbf{W}|\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_w)$ denote the probability of the observed passenger activity times conditional on the unknown parameters, and $\pi(\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_w)$ be the prior probability density of the parameters. Following Bayes' theorem, the posterior probability density of the unknown parameters given the observations is written as Eq. (11)

$$\pi(\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_w|\mathbf{W}) = \frac{f(\mathbf{W}|\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_w)\pi(\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_w)}{h(\mathbf{W})} \quad (4.11)$$

where $h(\mathbf{W})$ is the marginal likelihood also known as the model evidence. As is defined in Eq. (4.12), $h(\mathbf{W})$ is a normalizing constant in which all the unknown parameters are marginalized. It can only be presented by a high-dimensional integral with no closed form in this problem.

$$h(\mathbf{W}) = \int \int \int \int f(\mathbf{W}|\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W) \pi(\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W) d\mathbf{p}^g d\mathbf{p}^a d\alpha d\sigma_W \quad (4.12)$$

The posterior density is proportional to the numerator in Eq. (4.11) provided that the denominator is a normalizing constant. We then have

$$\pi(\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W | \mathbf{W}) \propto f(\mathbf{W}|\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W) \pi(\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W) \quad (4.13)$$

As the probability of observing passenger activity times $\mathbf{W}$ conditional on the unknown parameters is regarded as the likelihood of all the parameters given the observations, we have $f(\mathbf{W}|\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W) = L(\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W | \mathbf{W})$. We then can compute the total likelihood for the parameters by distinguishing the observed passenger activity time for each bus run k at each stop i as

$$L(\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W | \mathbf{W}) = \prod_{k=1}^K \prod_{i=1}^N f(W_{i,k} | \mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W) \quad (4.14)$$

As is stated in Eq. (4.3), we have

$$f(W_{i,k} | \mu_{i,k}, \sigma_W) = \frac{1}{\sqrt{2\pi\sigma_W^2}} \exp\left(-\frac{(W_{i,k} - \mu_{i,k})^2}{2\sigma_W^2}\right) \quad (4.15)$$

The probability of observing a specific passenger activity time $f(W_{i,k} | \mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W)$ can thus be rewritten as

$$f(W_{i,k} | \mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W) = f(W_{i,k} | \mu_{i,k}, \sigma_W) f(\mu_{i,k} | \mathbf{p}^g, \mathbf{p}^a, \alpha) \quad (4.16)$$

We next focus on $f(\mu_{i,k} | \mathbf{p}^g, \mathbf{p}^a, \alpha)$ which is the conditional probability of the expected passenger activity time given all the parameters of the gravity model. $\mu_{i,k} | Q_{i,j \in [i+1, N], k}, Q_{j \in [1, i-1], i, k}$ is deterministic as $f(Q_{i,j \in [i+1, N], k}, Q_{j \in [1, i-1], i, k})$ produces a single value for $\mu_{i,k}$ by using Eq. (4.2) and Eqs. (4.6) - (4.7). $f(\mu_{i,k} | \mathbf{p}^g, \mathbf{p}^a, \alpha)$ thus can be rewritten as Eq. (4.17).

$$f(\mu_{i,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha) = \prod_{j=i+1}^N f(Q_{i,j,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha) \prod_{j=1}^{i-1} f(Q_{j,i,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha) \quad (4.17)$$

To improve the interpretability of Eq. (4.17), we let $R(i)$ denote the set of possible OD pairs leaving from or destined to stop i . E.g. $R(3) = [(1,3), (2,3), (3,4), (3,5), \dots, (3,N)]$. We then rewrite Eq. (4.17) as

$$f(\mu_{i,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha) = \prod_{r \in R(i)} f(Q_{r,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha) \quad (4.18)$$

According to Eq. (4) and (5), we have two options to interpret $f(Q_{r,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha)$. Firstly we can take $Q_{r,k}$ as random variables that follow the Poisson distribution defined in Eq. (4.4). The conditional probability of observing $Q_{r,k}$ given the gravity parameters is obtained as

$$\begin{aligned} f(Q_{r,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha) &= f(Q_{r,k}|a_r \Delta_{*,k}) f(a_r \Delta_{*,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha) \\ &= \frac{(a_r \Delta_{*,k})^{Q_{r,k}} \exp^{-a_r \Delta_{*,k}}}{Q_{r,k}!} \end{aligned} \quad (4.19)$$

We can cancel $f(a_r \Delta_{*,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha)$ in that a_r is a definite product of the gravity model and $\Delta_{*,k}$ can be considered constant here. $\Delta_{*,k}$ always equals to the headway of bus run k at the origin stop of OD pair r .

Secondly, we can obtain $f(Q_{r,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha)$ using Eq. (4.5). In this case, $Q_{r,k}$ are not treated as the random variables, we thus can cancel $f(Q_{r,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha)$ in Eq. (4.19) as well as $f(\mu_{i,k}|\mathbf{p}^g, \mathbf{p}^a, \alpha)$ in Eq. (4.16).

Taken together, the total likelihood of the parameters is obtained as Eq. (4.20) if OD flows $Q_{r,k}$ are considered as random variables and are assumed to follow a Poisson distribution. Alternatively it is derived as Eq. (4.21) if $Q_{r,k}$ are considered as the expected OD flows in an observational period and the randomness in passenger' behavior is simplified. To focus on the mean estimation problem and reduce the computational complexity, we incorporate Eq. (4.21) in the Bayesian inference.

$$L(\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W | \mathbf{W}) = \prod_{k=1}^K \prod_{i=1}^N \left(f(W_{i,k} | \mu_{i,k}, \sigma_W) \prod_{r \in R(i)} f(Q_{r,k} | \mathbf{p}^g, \mathbf{p}^a, \alpha) \right) \quad (4.20)$$

$$L(\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W | \mathbf{W}) = \prod_{k=1}^K \prod_{i=1}^N f(W_{i,k} | \mu_{i,k}, \sigma_W) \quad (4.21)$$

By assuming that all the unknown parameters are independent and substituting Eq. (4.21) into Eq. (4.13), the posterior probability density can be rewritten as

$$\pi(\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W | \mathbf{W}) \propto \left(\prod_{k=1}^K \prod_{i=1}^N f(W_{i,k} | \mu_{i,k}, \sigma_W) \right) \pi(\mathbf{p}^g) \pi(\mathbf{p}^a) \pi(\alpha) \pi(\sigma_W) \quad (4.22)$$

It should be noted here that the elements in the parameter vector $\mathbf{p}^g$ and $\mathbf{p}^a$ are not perfectly independent in reality. We also need to specify the prior distribution for each unknown parameter to put Bayesian inference into practice. It is suggested in Sun et al. (2015) that a broad distribution such as uniform distribution should be used if there is very little prior knowledge available about the parameters. We thus assume uniform prior distribution for all the unknown parameters. The posterior distribution is expected to be corrected with the help of massive observations on passenger activity times in bus AVL data.

We also try to investigate the effect of additional observations on the estimation performance, making full use of the favorable flexibility of the Bayesian inference framework. For comparison purposes, we introduce passenger boarding and alighting counts into the observations and obtain the corresponding posterior density of the parameters. Let $\boldsymbol{\theta} = (\mathbf{p}^g, \mathbf{p}^a, \alpha)$ simplify the notation.

We firstly suppose that both boarding and alighting counts are available. In light of observed boarding and alighting flows, OD flows can be estimated without the information provided by the passenger activity time. Assume two normal distributions for $B_{i,k}$ and $A_{i,k}$ as in Eqs. (4.23) and (4.24).

$$B_{i,k} \sim N(\mu_{i,k}^B, \sigma_B^2) \quad (4.23)$$

$$A_{i,k} \sim N(\mu_{i,k}^A, \sigma_A^2) \quad (4.24)$$

Similar to Eq. (4.22), the posterior density for the parameters given the corresponding observations can be respectively derived as follows

$$\pi(\boldsymbol{\theta}, \sigma_B, \sigma_A | \mathbf{B}, \mathbf{A}) \propto \left(\prod_{k=1}^K \prod_{i=1}^N f(B_{i,k} | \mu_{i,k}^B, \sigma_B) f(A_{i,k} | \mu_{i,k}^A, \sigma_A) \right) \pi(\boldsymbol{\theta}) \pi(\sigma_B) \pi(\sigma_A) \quad (4.25)$$

$$\pi(\boldsymbol{\theta}, \sigma_B, \sigma_W | \mathbf{B}, \mathbf{W}) \propto \left(\prod_{k=1}^K \prod_{i=1}^N f(B_{i,k} | \mu_{i,k}^B, \sigma_B) f(W_{i,k} | \mu_{i,k}, \sigma_W) \right) \pi(\boldsymbol{\theta}) \pi(\sigma_B) \pi(\sigma_W) \quad (4.26)$$

$$\pi(\boldsymbol{\theta}, \sigma_A, \sigma_W | \mathbf{A}, \mathbf{W}) \propto \left(\prod_{k=1}^K \prod_{i=1}^N f(A_{i,k} | \mu_{i,k}^A, \sigma_A) f(W_{i,k} | \mu_{i,k}, \sigma_W) \right) \pi(\boldsymbol{\theta}) \pi(\sigma_A) \pi(\sigma_W) \quad (4.27)$$

where

$$f(B_{i,k} | \mu_{i,k}^B, \sigma_B) = \frac{1}{\sqrt{2\pi\sigma_B^2}} \exp\left(-\frac{(B_{i,k} - \mu_{i,k}^B)^2}{2\sigma_B^2}\right) \quad (4.28)$$

$$f(A_{i,k} | \mu_{i,k}^A, \sigma_A) = \frac{1}{\sqrt{2\pi\sigma_A^2}} \exp\left(-\frac{(A_{i,k} - \mu_{i,k}^A)^2}{2\sigma_A^2}\right) \quad (4.29)$$

4.5 Solution algorithms

As the normalizing constant is a high-dimensional integral with no-closed form, and conjugate prior is not available for the posterior density in Eq. (4.22) due to its complex formulation, it is difficult to analytically obtain the posterior distribution (a similar problem for Eqs. (4.25) - (4.27)). MCMC (Markov Chain Monte Carlo) methods are widely used to approximate the posterior distribution based on sampling candidate from a proposal distribution and updating the target posterior distribution. We implement MCMC methods here to draw the posterior distributions for the unknown parameters that

fit all the observed passenger activity times of each bus run at each stop. As the Bayesian posterior distribution is derived for each parameter (all estimated and derived parameters) by the MCMC methods, the posterior mean and confidence interval of OD flows, boarding/alighting flows as well as passenger loads are accordingly inferred. The robust estimation results are expected to relieve the operators of the uncertainties concerning actual passenger flows.

MCMC methods draw the posterior distribution for a high-dimensional parameter space and complicated model by constructing a Markov chain whose stationary distribution is forced to be equivalent or proportional to the posterior distribution of concern (Green and Worden, 2015). The sampling which is a process attaching values to the high-dimensional parameter vector is conducted iteratively. The attached values (current state) are only probabilistically dependent on the previous iteration (previous state). Metropolis-Hastings (M-H) algorithm is one of the most established MCMC methods (Metropolis et al, 1953; Hastings, 1970). Suppose $\boldsymbol{\delta} = (p_1^g, \dots, p_{N-1}^g, p_2^a, \dots, p_N^a, \alpha, \sigma_w)$, and let $\boldsymbol{\delta}^{(m)}$ denote the parameter vector at iteration m , also be the current state of the Markov chain. M-H algorithm proposes a new candidate state δ'_i for i th element of the parameter vector, which is generated from a proposal distribution $q(\delta'_i | \delta_i^{(m)})$ assumed to be symmetric and centered on $\delta_i^{(m)}$. The next step is to calculate the acceptance probability for δ'_i given by

$$\beta_{\text{M-H}}(\delta'_i, \delta_i^{(m)}) = \min \left\{ 1, \frac{\pi(\boldsymbol{\delta}' | \mathbf{W}) q(\delta_i^{(m)} | \delta'_i)}{\pi(\boldsymbol{\delta}^{(m)} | \mathbf{W}) q(\delta'_i | \delta_i^{(m)})} \right\} \quad (4.30)$$

where $\boldsymbol{\delta}'$ is identical to $\boldsymbol{\delta}^{(m)}$ except for the i th element.

We set $\delta_i^{(m)} = \delta'_i$ with probability $\beta_{\text{M-H}}(\delta'_i, \delta_i^{(m)})$ and repeat this process for each element of $\boldsymbol{\delta}^{(m)}$ in sequential order $i = 1, 2, \dots, 2N$. Next state $\boldsymbol{\delta}^{(m+1)}$ is obtained when the last element $\delta_{2N}^{(m)}$ is updated with probability $\beta_{\text{M-H}}(\delta'_{2N}, \delta_{2N}^{(m)})$. As $q(\delta'_i | \delta_i^{(m)})$ is required to be symmetric and centered on $\delta_i^{(m)}$, both Hazelton (2008) and Sun et al (2015) applied normal distribution and random walk for the candidate proposal, which simplifies the calculation in that $q(\delta_i^{(m)} | \delta'_i) = q(\delta'_i | \delta_i^{(m)})$ always holds. In addition, the acceptance probability cancels out the normalizing constant which is in the denominator of both $\pi(\boldsymbol{\delta}' | \mathbf{W})$ and $\pi(\boldsymbol{\delta}^{(m)} | \mathbf{W})$ and hence has an analytical solution given

$$\beta_{\text{M-H}}(\delta'_i, \delta_i^{(m)}) = \min \left\{ 1, \frac{f(\mathbf{W}|\delta')\pi(\delta')}{f(\mathbf{W}|\delta^{(m)})\pi(\delta^{(m)})} \right\} = \min \left\{ 1, \frac{f(\mathbf{W}|\delta')\pi(\delta'_i)}{f(\mathbf{W}|\delta^{(m)})\pi(\delta_i^{(m)})} \right\} \quad (4.31)$$

M-H algorithm is therefore considered well-suited for Bayesian inference. However, the aforementioned random walk proposal moves only in i th dimension at a time by fixing all the other elements of the parameter vector. It is not necessarily efficient in exploring a high-dimensional parameter space. Alternatively, the random walk proposal can move in all the dimensions at once, by sampling δ' conditional on $\delta^{(m)}$ and then conducting the accept-or-reject at last. This might vanish the acceptance probability when the dimension of the parameter space is high (Betancourt, 2017). Consequently, the proposal is always rejected and the Markov chain rarely moves.

Intuitively speaking, the random walk strategy might fail in high-dimension spaces where the number of directions to explore exponentially increases (Betancourt, 2017). Hamiltonian Monte Carlo (HMC) is a more efficient MCMC method than M-H algorithm in terms of sampling high-dimensional parameter spaces (Green and Worden, 2015; Betancourt, 2017), as it can maintain high acceptance rate even if it changes all the elements of the parameter vector simultaneously, and make large jumps into new unexplored regions.

Here the physical analogy is introduced for easier understanding. The state of a dynamical system can always be represented by a point (momentum, position) in phase space. Suppose each element of the parameter vector δ denote the position of a particle in one specific dimension, and introduce an auxiliary momentum $\mathbf{p}$ of the same size as δ , we have a dynamical system $(\mathbf{p}, \delta)$ whose joint probability density can be obtained as

$$\pi(\mathbf{p}, \delta) = \pi(\mathbf{p}|\delta)\pi(\delta|\mathbf{W}) \quad (4.32)$$

We write the density in terms of an invariant Hamiltonian function $H(\mathbf{p}, \delta)$, (Betancourt and Girolami, 2013; Betancourt, 2017).

$$\pi(\mathbf{p}, \delta) = \exp(-H(\mathbf{p}, \delta)) \quad (4.33)$$

The Hamiltonian of the system $(\mathbf{p}, \delta)$ can be written as

$$\begin{aligned}
H(\mathbf{p}, \boldsymbol{\delta}) &= -\ln \pi(\mathbf{p}, \boldsymbol{\delta}) \\
&= -\ln \pi(\mathbf{p}|\boldsymbol{\delta}) - \ln \pi(\boldsymbol{\delta}|\mathbf{W}) \\
&= T(\mathbf{p}, \boldsymbol{\delta}) + V(\boldsymbol{\delta})
\end{aligned} \tag{4.34}$$

where $T(\mathbf{p}, \boldsymbol{\delta}) = -\ln \pi(\mathbf{p}|\boldsymbol{\delta})$ is the kinetic energy and $V(\boldsymbol{\delta}) = -\ln \pi(\boldsymbol{\delta}|\mathbf{W})$ is the potential energy, and $H(\mathbf{p}, \boldsymbol{\delta})$ remains constant over the passage of time according to the law of conservation of energy. The potential energy is only dependent on $\pi(\boldsymbol{\delta}|\mathbf{W})$ which is also the target density of our Bayesian inference. By observing the system's evolving states over time, a trajectory passing through the high-dimensional space is derived. The trajectory contains a series of points constrained by $H(\mathbf{p}, \boldsymbol{\delta})$, which is thus equivalent to deriving a Markov chain whose stationary distribution is $\pi(\boldsymbol{\delta}|\mathbf{W})$ for the parameter vector $\boldsymbol{\delta}$. Time evolution of the system is defined by Hamilton's equations as

$$\begin{aligned}
\frac{d\boldsymbol{\delta}}{dt} &= +\frac{\partial H}{\partial \mathbf{p}} = \frac{\partial T}{\partial \mathbf{p}} \\
\frac{d\mathbf{p}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{\delta}} = -\frac{\partial T}{\partial \boldsymbol{\delta}} - \frac{\partial V}{\partial \boldsymbol{\delta}}
\end{aligned} \tag{4.35}$$

$\partial V/\partial \boldsymbol{\delta}$ is the gradient of the logarithm of the target density, indicating that HMC only works when gradient exists. Leapfrog integrator is usually employed to numerically approximate the new state $(\mathbf{p}', \boldsymbol{\delta}')$ from the current state $(\mathbf{p}^{(m)}, \boldsymbol{\delta}^{(m)})$. See Betancourt (2017) for more details. As we have $H(\mathbf{p}', \boldsymbol{\delta}') = H(\mathbf{p}^{(m)}, \boldsymbol{\delta}^{(m)})$, the new state always comes from the target density, which speeds up the convergence of the Markov chain and significantly improves the algorithm's efficiency. In practice, errors might be introduced by the leapfrog integrator, we cannot always accept the proposal. In the same way as in the M-H algorithm, the acceptance step is incorporated and the acceptance probability is written as Eq. (4.36), noting that β_{HMC} always equals to 1 in theory. A detailed discussion on correcting integrator errors can be found in Betancourt (2017).

$$\beta_{\text{HMC}} = \min \left\{ 1, \frac{\exp(-H(\mathbf{p}', \boldsymbol{\delta}'))}{\exp(-H(\mathbf{p}^{(m)}, \boldsymbol{\delta}^{(m)}))} \right\} \tag{4.36}$$

To keep the gradient, we rewrite the simultaneous form of the dwell time function by a smooth

maximum function as in Eq. (4.37).

$$\mu_{i,k} = \ln(\exp(t^b B_{i,k}) + \exp(t^a A_{i,k})) \quad (4.37)$$

We provide the algorithms using M-H as well as HMC as follows.

Input

Observations: $\mathbf{D}$, $\mathbf{W}$, Δ , K , N , t^b , t^a , c

Other input: number of chains, number of iterations (burn-in and sample size)

Output

Estimated parameters: $\mathbf{p}^g$, $\mathbf{p}^a$, α , σ_W

Derived parameters: $\mathbf{a}$, $\mathbf{Q}$, $\mathbf{B}$, $\mathbf{A}$, $\mathbf{O}$, $\boldsymbol{\mu}$

Initialization

Initialize random values for the parameter vector $\boldsymbol{\delta}^{(0)}$ for each chain

Set iteration counter $m \leftarrow 0$

Algorithm 1 M-H sampling

```

1 For each chain (can be done in parallel)
2   Repeat
3     For each element  $\delta_i^{(m)}$  ( $i = 1, \dots, 2N$ ) in the parameter vector  $\boldsymbol{\delta}^{(m)}$ 
4       Sample  $\delta'_i$  conditional on  $\delta_i^{(m)}$ , set  $\boldsymbol{\delta}' \leftarrow (\delta_1^{(m)}, \dots, \delta_{i-1}^{(m)}, \delta'_i, \delta_{i+1}^{(m)}, \dots, \delta_{2N}^{(m)})$ 
5       Compute  $f(\mathbf{W}|\boldsymbol{\delta}^{(m)})$ ,  $f(\mathbf{W}|\boldsymbol{\delta}')$  as Eq. (21) and  $\beta_{M-H}(\delta'_i, \delta_i^{(m)})$  as Eq. (4.31)
6       Set  $\delta_i^{(m)} \leftarrow \delta'_i$  with probability  $\beta_{M-H}(\delta'_i, \delta_i^{(m)})$ , otherwise  $\delta_i^{(m)} = \delta_i^{(m)}$ 
7     Set  $\boldsymbol{\delta}^{(m+1)} \leftarrow \boldsymbol{\delta}^{(m)}$  and  $m \leftarrow m + 1$ 
8   Until  $m =$  the number of iterations

```

Algorithm 2 HMC sampling

```

1 For each chain (can be done in parallel)
2   Repeat
3     Compute  $\boldsymbol{\delta}'$  given  $\boldsymbol{\delta}^{(m)}$  using Eq. (4.35) and leapfrog integrator
4     Compute  $\beta_{HMC}$  as Eq. (4.36)
5     Set  $\boldsymbol{\delta}^{(m+1)} \leftarrow \boldsymbol{\delta}'$  with probability  $\beta_{HMC}$ , otherwise  $\boldsymbol{\delta}^{(m+1)} = \boldsymbol{\delta}^{(m)}$ 
6     Set  $m \leftarrow m + 1$ 
7   Until  $m =$  the number of iterations

```

In each iteration m , a set of derived parameters can be computed using $\boldsymbol{\delta}^{(m)}$ by the following algorithm, which is also required when computing $f(\mathbf{W}|\boldsymbol{\delta}^{(m)})$

Algorithm 3 Obtaining derived parameters

```

1 For each bus stop  $i = 1, \dots, N$  and each bus stop  $j = 1, \dots, N$ 
2   Compute  $a_{i,j}$  as Eq. (4.9) based on  $\boldsymbol{\delta}^{(m)}$ 
3 For each bus run  $k = 1, \dots, K$  and each bus stop  $i = 1, \dots, N$ 
4   For each bus stop  $j = 1, \dots, N$ 
5     Compute  $Q_{i,j,k}$  as Eq. (4.5)
6   Compute  $B_{i,k}$  as Eq. (4.6),  $A_{i,k}$  as Eq. (4.7),  $O_{i,k}$  as Eq. (4.8) and  $\mu_{i,k}$  as Eq. (4.37)

```

In view of the high-dimensional parameter space in our problem, we select the HMC algorithm and use Stan which is a rapid sampler incorporating HMC algorithm and drawing multiple Markov chains in parallel to complete the parameter estimation. Further comparison between HMC and M-H is not discussed here and considered beyond the scope of this study.

Stan requires the users to fit the model input and output into 4 blocks: data, transformed data, parameter, transformed parameters. The items we classify for each block are as follows.

Data	$\mathbf{D}, \mathbf{W}, \Delta, K, N, t^b, t^a, c$
Transformed data	N/A
Parameter	$\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W$
Transformed parameter	$\alpha, \mathbf{Q}, \mathbf{B}, \mathbf{A}, \mathbf{O}, \mu$

4.6 Data processing

In the case study, the proposed methodology is tested on two main bus routes in Shizuoka City, Japan. Here we explain the data input to the model, then elaborate the case study results in Section 4.7.

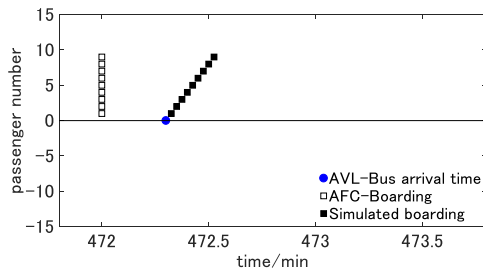
4.6.1 Dwell time, headway and vehicle activity time

The AVL data and AFC data in Shizuoka City, Japan are used in the experiment and the data period is 1-15, June 2016. 30 bus runs that depart from the initial stop during the morning peak 7:00 - 8:59 are selected in the sample dataset. Route A has 40 stops and Route B has 24 stops. There are thus 1200 observations of dwell times and headways for Route A and 720 observations for Route B.

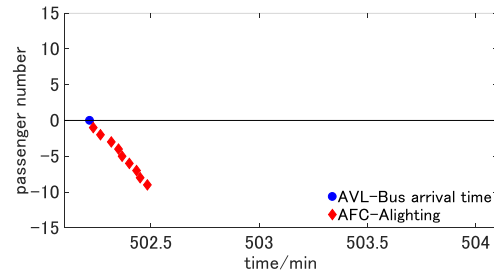
Dwell time and headway obtained from bus AVL data are the important model inputs. In the bus AVL data of Shizuoka City, the bus arrival time is based on the time stamp of door opening and is always recorded by the driver to keep complete bus trajectories even if there is no boarding or alighting passenger. The headway is obtained by using the bus arrival time. However, the departure time is not recorded in the bus AVL data, which generates the first difficulty in obtaining real bus dwell time. The departure time is then alternatively inferred by using the AFC data to obtain the time stamp of the last transaction at each bus stop, and the time interval between the bus arrival time obtained from AVL

data and the last transaction obtained from AFC data is taken as the dwell time. Therefore, the vehicle activity time is simplified to zero in the data input to the model.

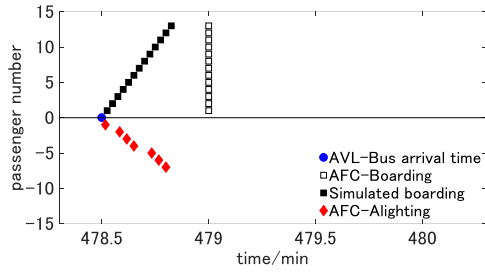
The second difficulty is that the time stamp of boarding transactions is always rounded by a full minute in the AFC data, e.g. the boarding transactions occur from 7:01:00 to 7:01:59 are all rounded as 7:01:00. A list of estimated average boarding time in existing literature can be found in Tirachini (2013). The average boarding time varies widely from 1.64sec to 8.87sec for the passengers using different payments or boarding buses with steps at the entrance. Sun et al (2014) estimated 1.95 - 2.05 for the value of card user's average boarding time after excluding the effect of total boarding number, in-vehicle crowding and steps at the entrance. We then assume a constant boarding interval of 1.5sec for each passenger to derive the total boarding time and the last boarding transaction. It is slightly shorter than the findings in the above literature, as we intend to limit the effect of simulated data on the synthesized data. And we assume boarding activity and alighting activity always proceed simultaneously. This problem appears important in the cases that the simulated total boarding time is longer than the total alighting time. Four patterns of transaction dynamics can be observed from the AFC data as is shown in Figure 4.2. The departure time, as well as the dwell time in (a) and (c), has to be determined by the total boarding time which is longer than the total alighting time but is simulated. The inherent randomness in dwell time is thus eliminated in these cases. For Case (d), the randomness is also slightly reduced, as the departure time might be determined by the boarding time even if there are few boarding passengers, due to in-vehicle overcrowding, wheelchair moving, etc. 422 out of the 1200 dwell times of Route A and 315 out of 720 dwell times of Route B are determined by the simulated total boarding time.



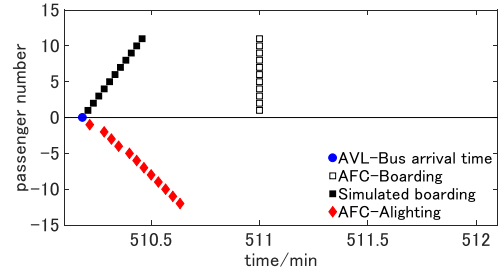
(a) Only boarding passengers



(b) Only alighting passengers



(c) Both boarding and alighting passengers, the simulated total boarding time is longer than the total alighting time



(d) Both passengers exist, the simulated total boarding time is shorter than the total alighting time

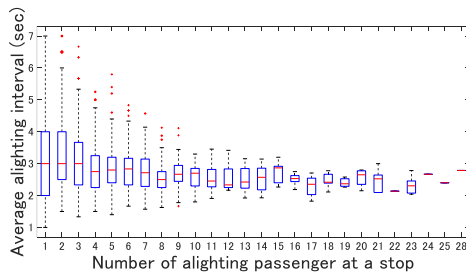
Figure 4.2 Patterns of boarding and alighting transactions observed from AFC data

4.6.2 Boarding/alighting time per passenger

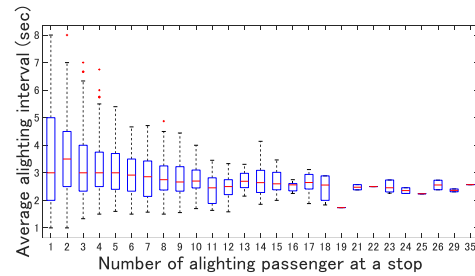
In addition to dwell time and headway, boarding/alighting time per passenger is also the data input to the model. Boarding time per passenger is assumed as 1.5sec for the reasons explained above, and alighting time per passenger is obtained by a regression analysis using the AFC data. As is shown in Table 4.1 and Figure 4.3, the average alighting time per passenger is 2.5sec. In addition, Figure 4.3 illustrates that the average alighting time varies significantly when there are only a few alighting passengers, justifying our assumption in Eq. (4.3), and indicating that the standard deviation σ_W is not necessarily in proportion to the expected passenger activity time which is positively correlated to the passenger numbers.

Table 4.1 Regression analysis results of alighting time per passenger

	t^a	Intercept	R ²
Route A	2.5008	1.3392	0.8664
Route B	2.5162	1.8068	0.8715



(a) Route A



(b) Route B

Figure 4.3 Average alighting time against different number of alighting passengers

4.7 Estimation results

The case study is designed to illustrate

- (1) The estimation performance of using bus AVL data only, regarding the aggregated boarding/alighting flows, and the disaggregated OD flows.
- (2) The effect of additional observation on the estimation performance, e.g. combining passenger boarding/alighting counts with bus AVL data.
- (3) The effect of different prior knowledge concerning the unknown parameters on the estimation performance, e.g. setting different bounds for the uniform distribution applied on the unknown parameters.

Serving for the above three objectives, six scenarios are distinguished.

Scenario 1 (AB): unchained boarding-alighting counts are added to the observations to substitute for the bus dwell times. $\delta = (\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_A = \sigma_B)$. $f(\mathbf{B}, \mathbf{A}|\delta)$ is computed as Eq. (25).

Scenario 2 (BT): boarding counts and bus dwell times are used as the observations. $\delta = (\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W, \sigma_B)$. $f(\mathbf{B}, \mathbf{W}|\delta)$ is computed as Eq. (26).

Scenario 3 (AT): alighting counts and bus dwell times are used as the observations. $\delta = (\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W, \sigma_A)$. $f(\mathbf{A}, \mathbf{W}|\delta)$ is computed as Eq. (27).

Scenario 4 (AT*): the same observations as AT are used, and more detailed prior information is given on the bounds of unknown parameters. $\delta = (\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W, \sigma_A)$. $f(\mathbf{A}, \mathbf{W}|\delta)$ is computed as Eq. (27).

Scenario 5 (T): only bus dwell times are used as the observations. $\delta = (\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W)$. $f(\mathbf{W}|\delta)$ is computed as Eq. (21).

Scenario 6 (T*): the same observations as T are used, and more detailed prior information is given on the bounds of unknown parameters. $\delta = (\mathbf{p}^g, \mathbf{p}^a, \alpha, \sigma_W)$. $f(\mathbf{W}|\delta)$ is computed as Eq. (21).

Table 4.2 lists the elapsed time of running the HMC algorithm via Stan for each scenario. The sampling for each chain is conducted in parallel. For both routes, the scenarios using dwell times as the main observations cost the least time (T* for Route A, T for Route B). Although only one more parameter

is introduced, the elapsed time in BT, AT and AT* is 1.2 times as that in T for Route A and 1.5 times for Route B. The effect of specifying bounds for the unknown parameters on the elapsed time is inconclusive. The time cost in AT* (T*) is not necessarily longer or shorter than that in AT (T).

Table 4.2 Elapsed time of each scenario

	#Run	#Stop	#Parameter	#Chain	#Iteration	Elapsed time (sec)
Route A						
AB	30	40	80	3	4000	651.45
BT	30	40	81	3	4000	627.91
AT	30	40	81	3	4000	828.54
AT*	30	40	81	3	4000	754.07
T	30	40	80	3	4000	606.99
T*	30	40	80	3	4000	538.72
Route B						
AB	30	24	48	3	4000	207.99
BT	30	24	49	3	4000	243.38
AT	30	24	49	3	4000	244.12
AT*	30	24	49	3	4000	268.49
T	30	24	48	3	4000	164.10
T*	30	24	48	3	4000	175.55

The results are based on a MacBook Pro 2017 with 2.5GHz Intel Core i7 processor and 16GB 2133 MHz LPDDR3 memory, using CmdStan 2.17.1.

4.7.1 Estimated unknown parameters

Prior knowledge is critical for deriving reliable estimation result in Bayesian inference. Initially, we assume uniform distribution and apply rough bounds $U[0,10]$ on each element of $\mathbf{p}^g$ and $\mathbf{p}^a$ in Scenario AB, BT, AT and T. For stop i at which no passenger boards or alights in the observational period (30 bus runs), we set $U[0,0]$ for the corresponding p_i^g or p_i^a .

As is illustrated in Fig. 4.4, Scenario AB and BT derive similar posterior means for $\mathbf{p}^g$ which however are significantly different from those produced by AT and T. The distinction among these four scenarios with respect to $\mathbf{p}^a$ is much less obvious. We find that more accurate posterior mean and confidence interval are derived by AB and BT, by comparing the derived passenger boarding/alighting flows, passenger loads and OD flows of all the four scenarios. The estimated

boarding flows and OD flows by AT and T are in particular deviating from the ground truth, see Fig. 4.6 for details. Therefore, we conclude that it is difficult to conduct reliable inference by merely employing dwell time observations and vague prior information on the unknown parameters. Solid results are unsurprisingly obtained by Scenario AB, as boarding and alighting counts contain abundant information regarding the generation and attraction power of the bus stops. Furthermore, the outperformance of boarding counts over alighting counts in terms of playing the role of observations in this inference problem is an interesting discovery. We thus suggest that the transit operator with a limited budget can install the sensors that collect boarding count data in priority.

In order to test the effect of prior information, we leverage the posterior estimates derived by Scenario AB to enrich the prior knowledge we have on the unknown parameters. In Scenario AT* and T*, we specify the bounds for stop generation power roughly according to, albeit not precisely as the posterior confidence interval inferred by Scenario AB and leave stop attraction power unconstrained. Specific bounds are imposed on the stops with explicit characteristics. $U[9,10]$ is specified for the railway station (Stop 27) in the middle of Route A, while $U[0,4]$ for a series of stops (Stop 13 - 18) across the industrial area. Also some stops are constrained simply following the results of AB, e.g. $U[0,1]$ for many stops (Stop 1 - 25) without particular landmarks on Route A, $U[9,10]$ for Stop 19 which is named by harbor bridge and does not intuitively relate to high trip generation power. Table 4.3 summarizes the various prior information provided for p^g and p^a in different scenarios. Prior (Prior*) specifies the lower and upper bounds for p^g and p^a in AB, BT, AT and T (AT* and T*). By imposing stricter constraints on p^g for a few stops, the posterior means of p^g in AT* and T* naturally move toward those in AB and BT, as is shown in Fig. 4.4. The relatively insignificant gap between AB/BT and AT*/T* regarding p^a becomes even closer for both routes.

Table 4.3 Different prior information on p^g and p^a

		Prior (AB, BT, AT, T)	Prior* (AT*, T*)	Stop characteristics
Route A	Bus stop			
$p_i^g, i =$	1-25	$U[0,10]$	$U[0,1]$	Stops without particular landmarks
	26	$U[0,10]$	$U[1,2]$	City hall
	27	$U[0,10]$	$U[9,10]$	Railway station
	28-34	$U[0,10]$	$U[0,1]$	Stops without particular landmarks
	35-39	$U[0,0]$	$U[0,0]$	End of the route
$p_i^a, i =$	2-13	$U[0,0]$	$U[0,0]$	Start of the route
	14-40	$U[0,10]$	$U[0,10]$	
Route B	Bus stop			
$p_i^g, i =$	1-12	$U[0,10]$	$U[0,10]$	
	13-18	$U[0,10]$	$U[0,4]$	Mostly industrial area
	19	$U[0,10]$	$U[9,10]$	Harbor bridge
	20-22	$U[0,10]$	$U[0,10]$	
	23	$U[0,0]$	$U[0,0]$	End of the route
$p_i^a, i =$	2-11	$U[0,0]$	$U[0,0]$	Start of the route
	12-24	$U[0,10]$	$U[0,10]$	

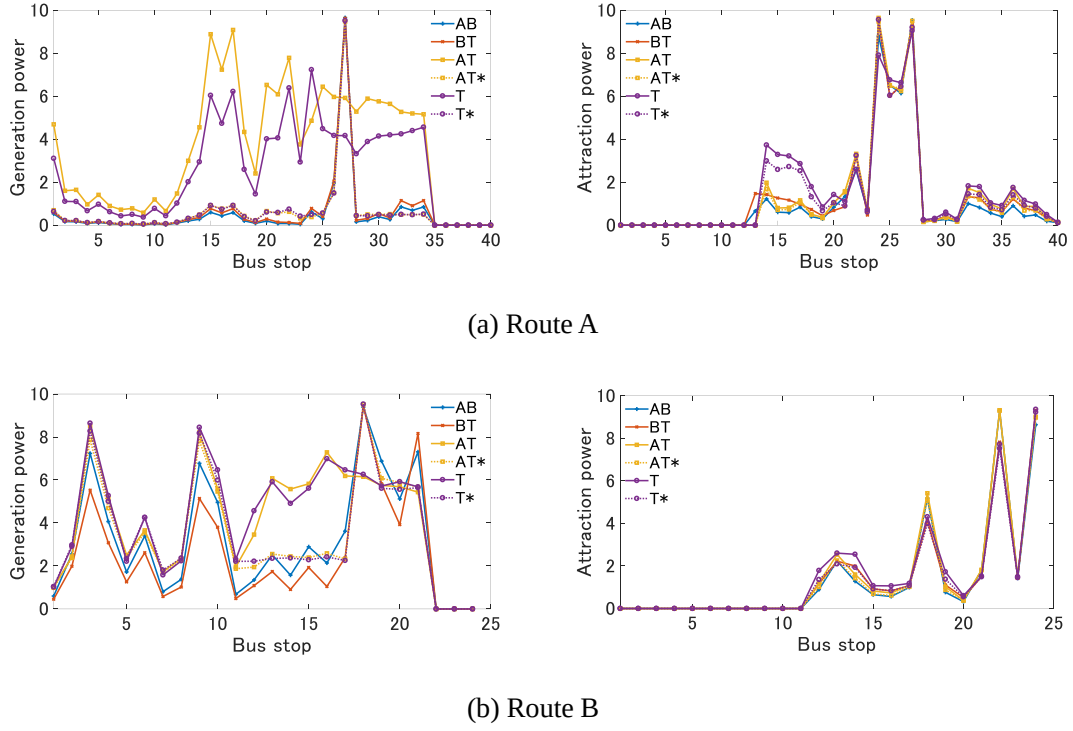
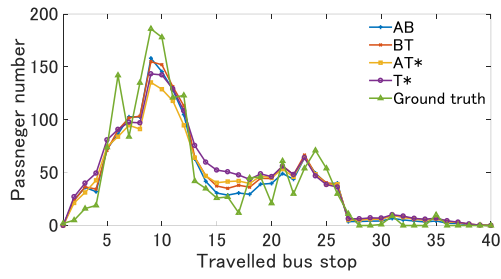


Figure 4.4 Posterior means of p^g and p^a

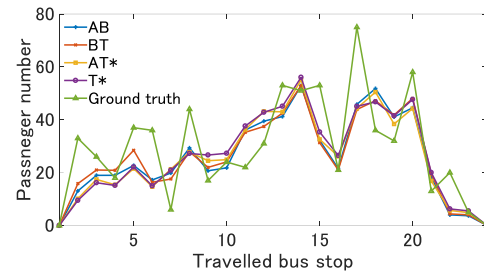
Hereinafter we narrow the focus of the comparison to AB, BT, AT* and T* only. As is shown in Table 4.4, we assume identical and ambiguous prior distribution for α , σ_W , σ_B and σ_A , and derive similar posterior means for σ_W , σ_B and σ_A in these four cases. The posterior means of α vary in each case due to the different bounds applied and the resulting p^g and p^a . To investigate the model fit of the modified gravity model, the real and inferred distributions of passenger in-vehicle time in terms of travelled bus stops are illustrated in Figure 4.5. The modified gravity model successfully captures the travel pattern although the estimated pattern is smoother than the real pattern and overestimation or underestimation is inevitably generated. We also note that the incorporated gravity model is not able to replicate the significant sudden changes. It is observed on Route A that the sharp rise in passenger numbers from traveling five stops to six stops is followed by a steep drop from six to seven stops. Some dominant OD pairs exist on the transit route, e.g. Stop 27 (transit hub) to Stop 33 (industrial area) on Route A, and the corresponding travelers might be indifferent on the distance cost. This limitation of the gravity model is supposed to account for some estimation errors of passenger flows.

Table 4.4 Prior information and posterior means of α , σ_W , σ_B and σ_A

	Prior	Posterior mean			
		AB	BT	AT*	T*
α	$U[0,10000]$	418.1847	628.7502	879.7225	866.8917
σ_W	$U[0,60]$	N/A	4.0876	4.1266	4.0974
σ_B	$U[0,10]$	1.2187	1.2612	N/A	N/A
σ_A	$U[0,10]$	1.2187	N/A	1.1785	N/A

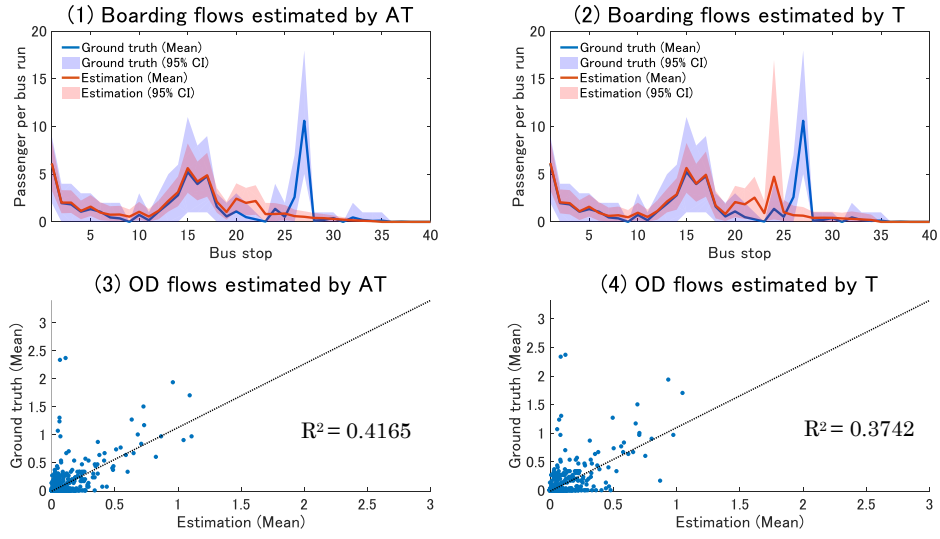


(a) Route A

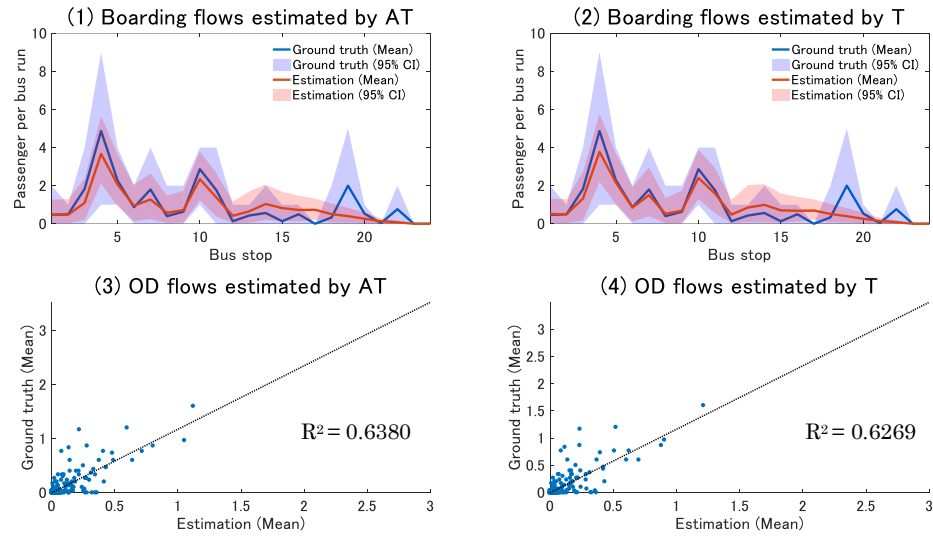


(b) Route B

Figure 4.5 In-vehicle time distribution



(a) Route A



(b) Route B

Figure 4.6 Estimation results of boarding flows and OD flows by Scenario AT and T

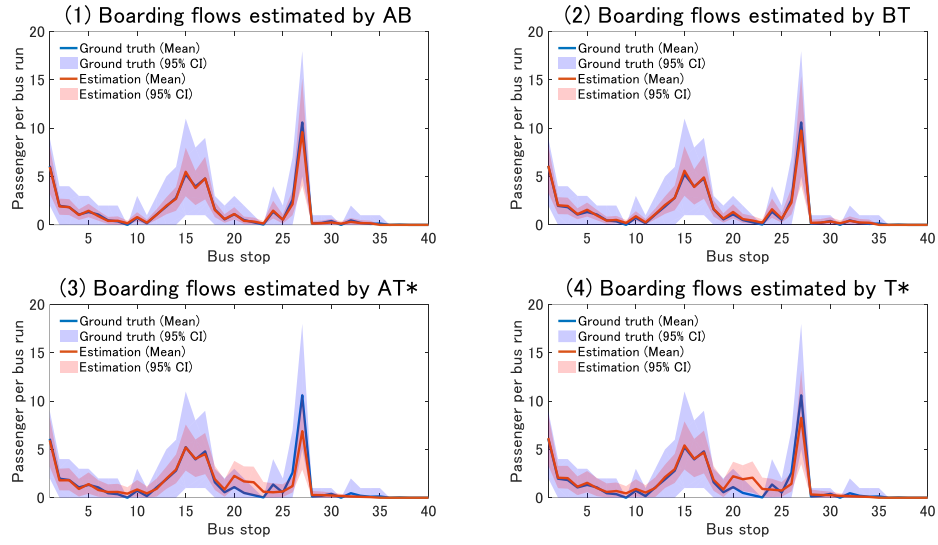
4.7.2 Estimated passenger boarding and alighting flows, as well as passenger loads

Figures 4.7 – 4.9 illustrate the posterior mean and 95% confidence interval (CI) of the estimated boarding flows, alighting flows and passenger loads. The ground truth is the real numbers of boarding (alighting, on-board) passengers at each stop observed from 30 bus runs leaving the initial stop during 7 am – 8 am. We demonstrate the mean value and the variation of the 30 bus runs. The estimation objectives are thus also the mean and the variation. The estimated CI is narrower than the real one, as

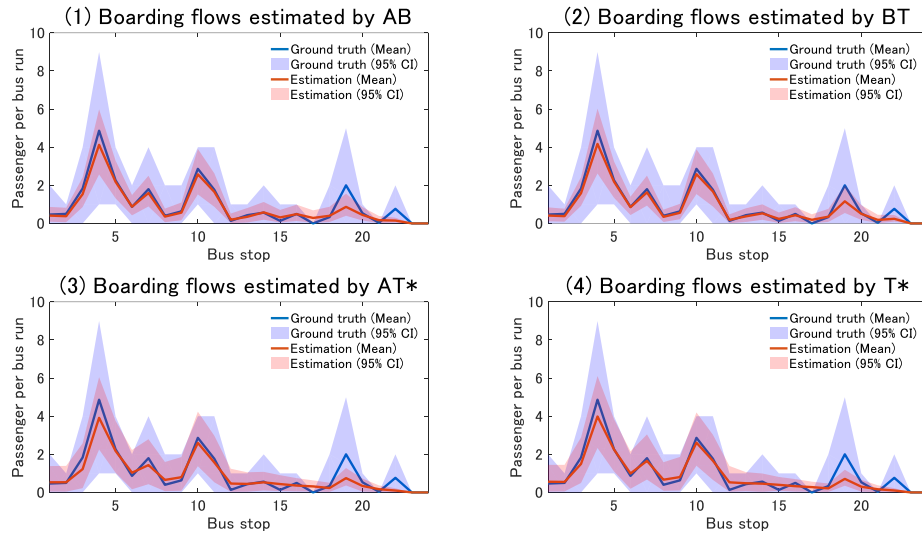
we assume that the variation in passenger flows only results from the variation in headways in order to reduce the computational complexity. This assumption is not realistic in that the inherent randomness in passenger behavior might overwhelm the variation in headway, and become the main cause of the fluctuation in passenger flows. Nevertheless, these figures convey a same favorable message that the mean passenger flows (loads) along the bus route are accurately estimated, and the variations in an observational period are reliably inferred in Scenario AB and BT, as well as in Scenario AT* and T* if rich prior information is provided. The robustness of methodology is verified by Route A and Route B which have fairly different route configurations. Route A has a transit hub in the middle, and the final stop of Route B is a railway station. RMSE and MAPE are listed in Table 4.5. They are based on the mean values as we do not consider reproducing the flows (loads) for each bus run. We highlight the worst scenario regarding each evaluation index for each evaluation index, and we find that the worst ones are not necessarily produced by Scenario T*, although it is fed with least direct observations on passenger flows. It outperforms AT* in terms of boarding flows and matches the performance of BT* with respect to alighting flows. Regarding passenger loads, it underperforms all the other cases for Route A though, the downward deviation is acceptable. We also notice that it generates the lowest error rates for Route B's passenger loads, but it might result from errors of boarding and alighting flows occasionally cancelling out each other.

Table 4.5 RMSE/MAPE of estimated passenger boarding/alighting flows and passenger loads

	Boarding flows		Alighting flows		Passenger loads	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Route A						
AB	0.1964	7.60%	0.1421	7.36%	1.0964	4.36%
BT	0.1903	8.97%	0.3995	19.05%	1.3421	5.57%
AT*	0.7466	26.65%	0.1663	7.72%	1.6106	6.10%
T*	0.6230	25.46%	0.4075	20.48%	2.5300	10.45%
Route B						
AB	0.3306	19.87%	0.3188	15.09%	1.5624	11.25%
BT	0.2703	16.71%	0.5007	21.92%	1.6020	11.02%
AT*	0.4109	28.10%	0.3316	15.38%	1.3830	9.99%
T*	0.3902	26.00%	0.4956	22.41%	1.0115	6.73%

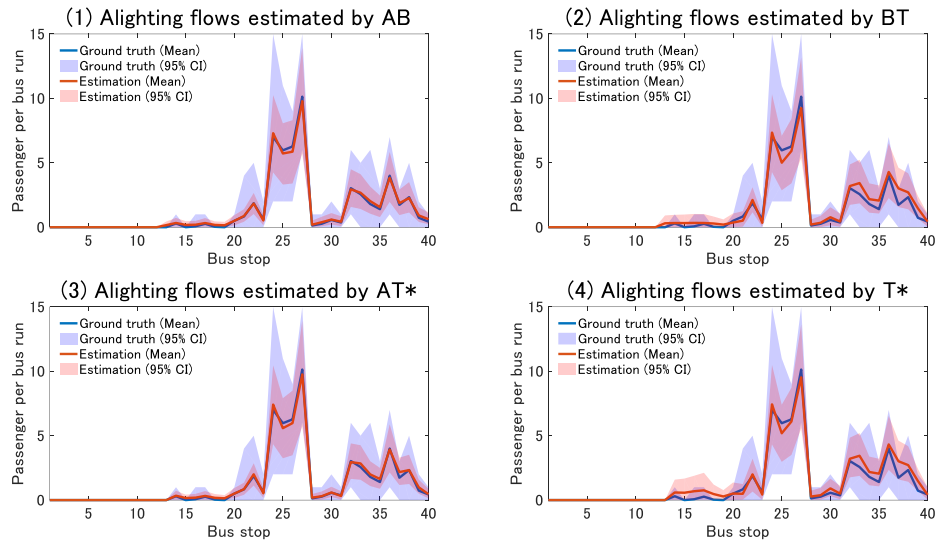


(a) Route A

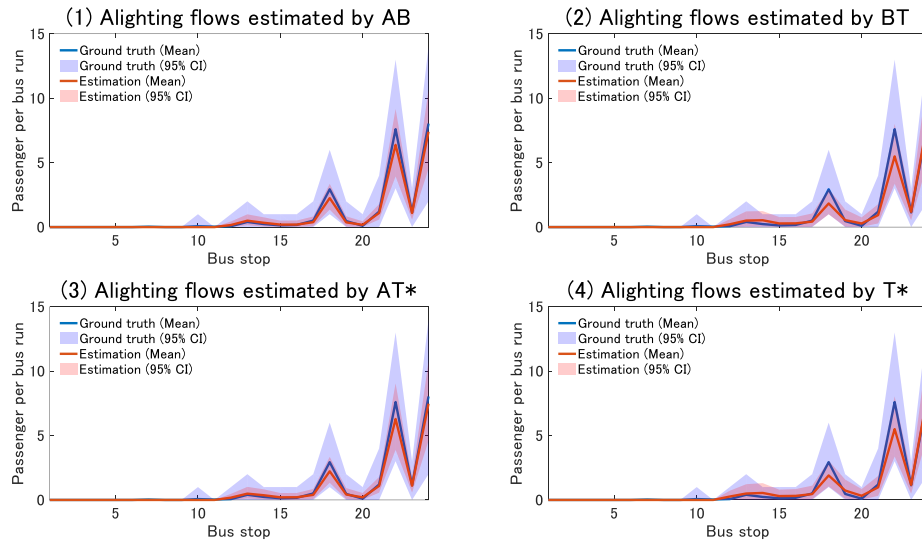


(b) Route B

Figure 4.7 Estimated passenger boarding flows

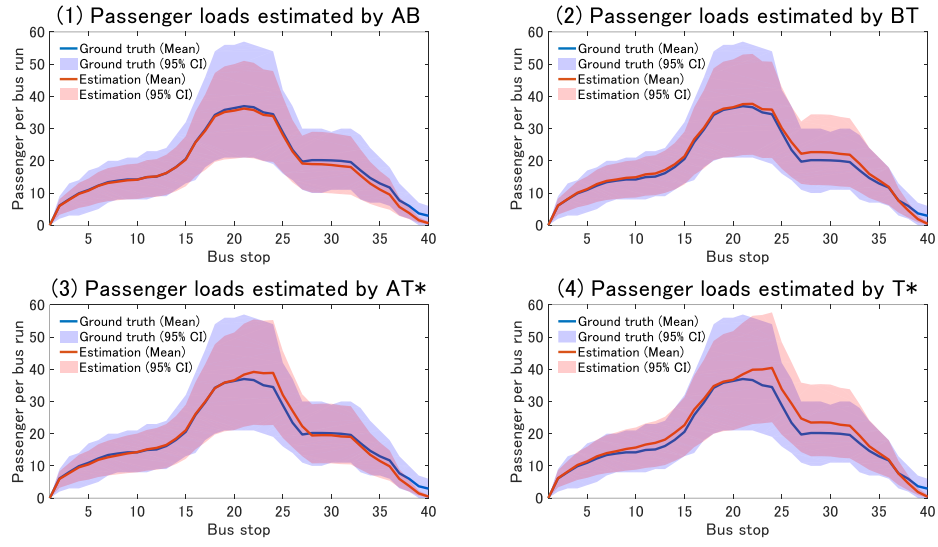


(a) Route A

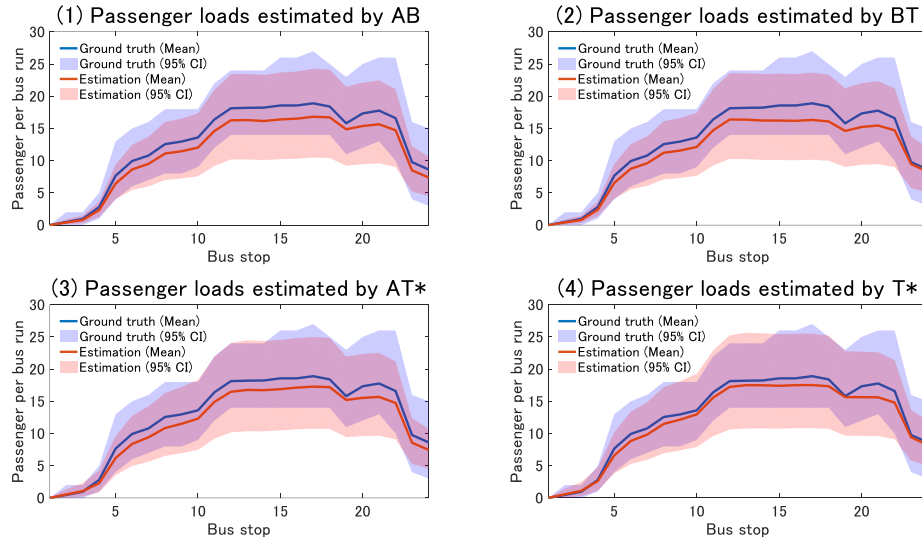


(b) Route B

Figure 4.8 Estimated passenger alighting flows



(a) Route A



(b) Route B

Figure 4.9 Estimated passenger loads

4.7.3 Estimated passenger OD flows

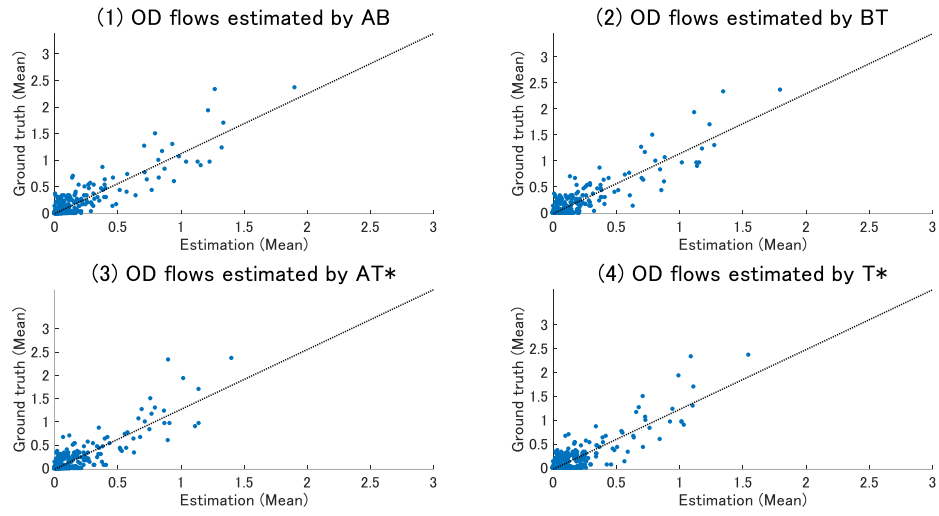
We then discuss the estimation performance of OD flows which is the estimation objective of most concern. Regression analysis is conduct for each scenario to evaluate the estimation performance, using ground truth as the dependent variable and estimated mean as the independent variable. The OD pairs having zero flow are excluded from the analysis. Fig. 10 indicates the model fit. The regression

analysis estimates and error rates can be found in Table 4.6. There is no obvious gap in the performance between T^* and the other three cases, while we acknowledge the slight underperformance of T^* compared with AB and BT. Figure 4.10 shows that large flows are accurately estimated in all the scenarios. We parameterize the OD estimation problem by the arrival rates per minute. This assumption is more suitable in capturing regular trips that are observed on most of the bus runs than random ones which are rarely observed on few bus runs. The general estimation accuracy in the form of R^2 ranges from 76% to 82% for Route A, from 70% to 75% for Route B. Although T^* produces the worst performance for both routes, the 5% - 6% downward deviation from the performance of AB is considered acceptable given such limited observations as merely bus dwell times and inferred prior information. The calculation of MAPE is based on the average flow over all the OD pairs to avoid introducing a close-to-zero number in the denominator. Nevertheless, we still derive inflated MAPEs due to the sparseness of the upper triangle OD matrix.

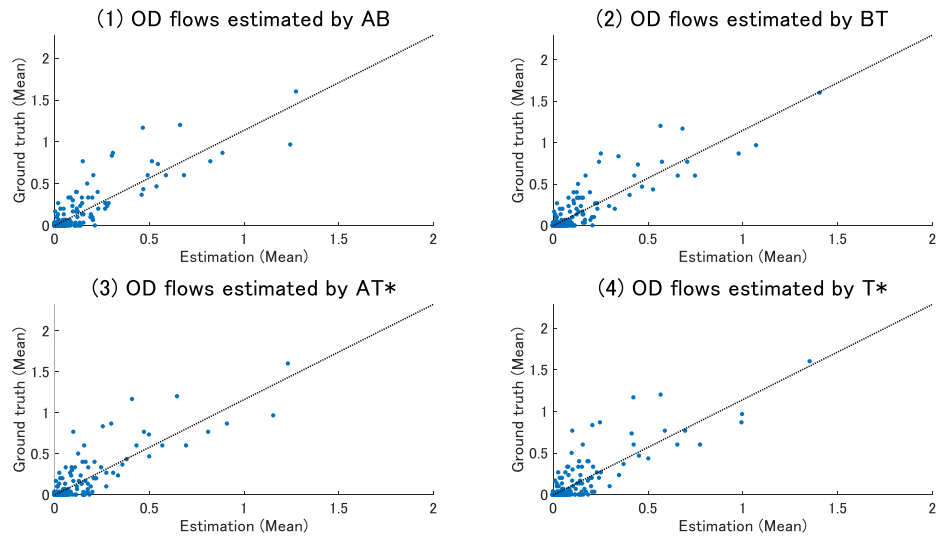
Figure 4.11 investigates into the estimation errors of T^* by presenting the errors in a matrix. In addition, it highlights the stop with the largest errors and illustrates the gap between estimated and real boarding/alighting flows at this stop. It also demonstrates the noteworthy strength of the proposed method that the confidence interval is obtained for each OD flow departing from or destined to a specific stop. The derived mean and 95% CI generally captures the expected OD flows and part of the actual variation which is implied by 75% CI due to overwhelming zero observations of OD flows. We note that the CI estimated by this method is the probable variation of the mean value. An additional step to draw the unknown parameters from the posterior distribution combined with sampling the OD flows from Poisson distribution to match the specific dwell time/passenger activity time is not conducted so that the random zero or large flows are frequently out of the estimated range.

Table 4.6 Model fit and RMSE/MAPE of estimated passenger OD flows

	Beta	Intercept	R ²	RMSE	MAPE
Route A					
AB	1.1297	-0.0077	0.8208	0.1028	62.16%
BT	1.1545	-0.0146	0.8100	0.1067	68.77%
AT*	1.2858	-0.0198	0.7613	0.1240	75.65%
T*	1.2529	-0.0263	0.7621	0.1224	81.51%
Route B					
AB	1.1379	0.0014	0.7539	0.1151	58.52%
BT	1.1494	-0.0007	0.7721	0.1114	59.35%
AT*	1.1634	-0.0012	0.7231	0.1224	61.86%
T*	1.1450	-0.0023	0.7051	0.1253	64.66%



(a) Route A



(b) Route B

Figure 4.10 Estimated passenger OD flows

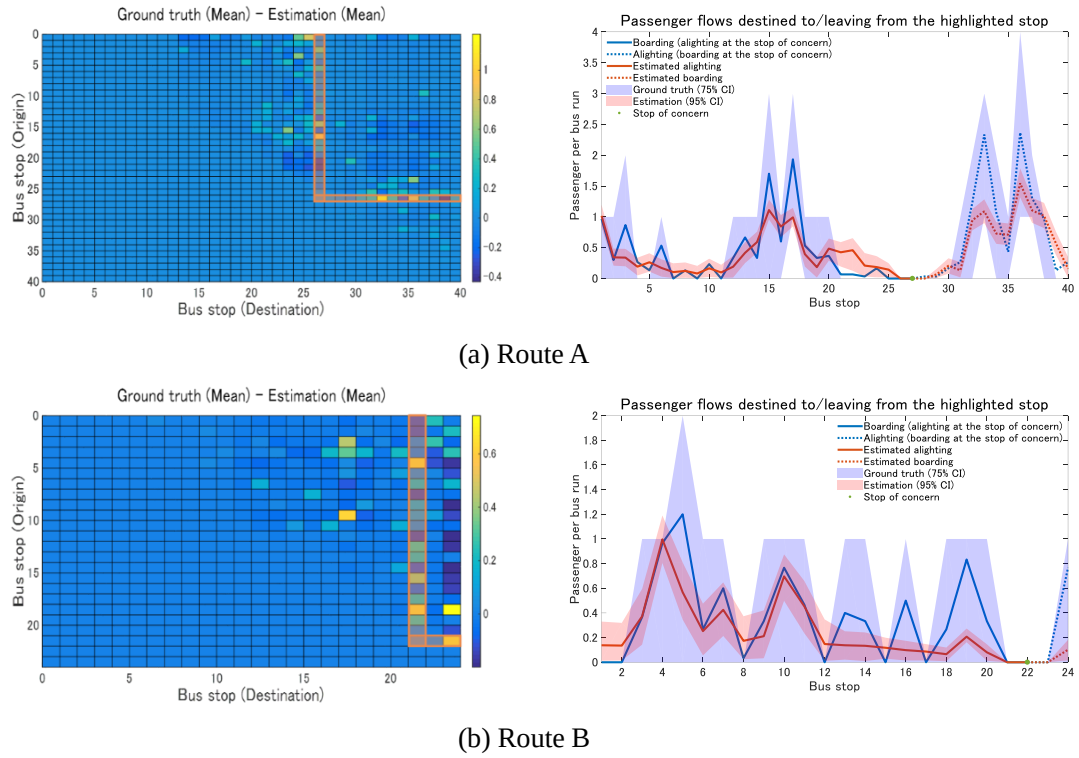


Figure 4.11 Estimation errors of passenger OD flows by T*; Left: OD matrix errors highlighting the stop with largest errors, Right: Passenger flows destined to/leaving from the highlighted stop

4.8 Findings

This chapter proposes a novel methodology to estimate boarding/alighting flows, passenger loads and OD flows for a transit route, using bus AVL data as the main data source. We extract bus dwell times and headways from the bus AVL dataset. We focus on the expected OD flow rates which are further explained by a modified gravity model using the generation and attraction power of the stops as the unknown parameters. Bayesian inference is applied to estimate the unknown parameters conditional on all the observed bus dwell times.

The methodology is tested on two bus routes in Shizuoka City, Japan, and the estimation performance is validated by AFC data. We distinguish six scenarios to explore the potential of dwell times in estimating passenger OD flows and additional information that can enhance the estimation performance. It is found difficult to correctly estimate the OD flows if only dwell times are observed. Accurate posterior means and reliable posterior confidence intervals are derived for passenger boarding/alighting flows, passenger loads and OD flows, with the help of detailed prior information on the unknown parameters which are inferred by using passenger boarding and alighting count data.

We also find that boarding counts are more helpful supplement data than alighting counts. In light of the reliable estimation results obtained by the proposed methodology, the operators having no direct observation on passenger flows are now capable of delivering high-quality services that can favorably accommodate the fluctuated actual user demands.

4.9 Limitations and extensions

Several extensions that can potentially strengthen the contribution of this research should be noted. Firstly, the method to collect the prior information on stop importance remains implicit. In the case study, we demonstrate that more precise prior information in terms of the bounds on the uniform distribution assumed for the unknown parameters can greatly improve the estimation performance when only dwell times are employed as the observation. However, we acknowledge that we lend the hand of passenger count data and it is difficult to directly observe the lower and upper bounds of the stop importance in the real world. The data sources that are conventionally used to calibrate a gravity model, such as population data, are not easy to collect at a bus stop base. We thus provide some candidate data here that have the potential to explain the generation and attraction power of the stop, and can be collected at a low cost. Number and property of the POIs (Points of Interest) within a certain distance from the bus stop are considered powerful in explaining the importance of stops, and they can be freely obtained from Open Street Map or other map databases. In addition, the topology of the bus transit network and the connectivity of the bus stop are strong supplements. With these data, it is also expected to calibrate a gravity model considering the weights of different types of POIs contributing to the generation and attraction power, as well as network-based cost function.

Secondly, advanced dwell time models that can account for complex scenarios are expected to further enhance the transferability of the proposed methodology. Bus bunching and in-vehicle crowding might significantly prolong the bus dwell times. Without a dwell time model specifically addressing these problems, the passenger flows are likely to be overestimated. However, the parameters are concerned not to converge if too many parameters are introduced or the model structure is highly complicated in such an underspecified problem. The balance between dwell time structure and computational complexity is a worthy discussion.

Last but not least, other interesting extensions include reconstruction of the OD matrix for each bus

run and prediction on the passenger loads in real time, given the estimated mean OD flow rates and bus AVL data. A two-stage Bayesian inference approach based on the proposed methodology is a feasible solution.

Chapter 5 **Complex dwell time models**

5.1 Introduction

In Chapter 4, an estimation methodology is provided to infer the OD matrix from bus AVL data. One important limitation is that the dwell time model employed is a basic one, without considerations on in-vehicle crowding, the effect of different methods of payment, and the effect of bus-to-bus interaction such as bus bunching on the bus dwell time. In-vehicle crowding and bus bunching may greatly extend the bus dwell time, and more importantly passenger boarding and alighting activities sometimes do not proceed in the extended dwell time. Here we consider the bus has to wait in a bus queue in the dwelling process, and only one bus is allowed to load passengers in the bus berth at the same time. Consequently, the buses except for the first bus in the queue have to unload the alighting passengers first, wait until the leading bus departs, and then move to the berth to load passenger. There is a dead time depending on the length of the queue and the dwell times of the previous buses. In this case, the total bus dwell time for the back bus sequentially consists of vehicle activity time (arriving), passenger alighting time, waiting time in the queue, vehicle activity time (moving to the berth), passenger boarding time and finally vehicle activity time (leaving). Significant overestimation can be produced if passenger activities are considered to proceed during the intermediate vehicle waiting and activity time. Multiple stopping and moving forward is possibly required for a bus at the same stop if more than two buses are queueing. In addition, in-vehicle crowding may delay the boarding process. Boarding passengers may wait first until some space is released by alighting passengers. We also consider the different boarding/alighting time per person of various methods of payment for the dwell time model in this chapter.

In order to collect the data required to model the effect of (1) bus-to-bus interactions; (2) in-vehicle crowding; (3) methods of payment on the bus dwell time, we conduct an onboard survey in Kyoto City. The fitted dwell time models are incorporated into the OD estimation framework proposed in Chapter 4. The case study in this chapter then is based on a bus route frequently faced with bunching (both bunched with the buses of the same line and those from other lines) and crowding.

5.2 Bus dwelling process

TCRP Report 19 (1996) defines two types of bus stops: curbside stops and bus bays. A curbside bus stop usually only requires a sign to designate a stop, it is thus convenient to install and relocate. It also

imposes more exposure of road traffic to the bus vehicle, possibly preventing lane changes to avoid delay behind stopped buses (Meng and Qu, 2013). Bus bay is a deliberately constructed area comprising vehicle entry, passenger boarding/alighting and vehicle exit segments, separating the bus stop from road traffic to some degree. As a cost, the bus has to wait at the exit area until there is enough space to insert back to the road traffic.

Meng and Qu (2013) estimate the distribution of time required to merge into the road traffic for bus bays. Bian et al. (2015) on the other hand focus on the bus queueing time at a curbside bus stop. They assume a bus stop with two berths that can load/unload the passengers simultaneously and apply no-overtaking rule to ease the model complexity. Dai et al. (2019) model the trip travel time for a bus line with an explicit consideration on the complete process of bus pre-loading/unloading to post-loading/unloading. They decompose the bus dwelling process into entering the bus stop (pre-loading/unloading stage), loading/unloading passenger and merging into the road traffic (post-loading/unloading stage). They also consider the extra waiting time due to queueing for entering the passenger boarding/alighting area in a no-overtaking situation (first-in-first-out queue). Taken together, bus dwell time therefore is a production of the waiting time in the bus queue, passenger activity time and time required to merge into the road traffic. Dai et al. (2019) provide a good illustration of the whole process, as is shown in Figure 5.1. Omitting the indexes of bus run and bus stop, let T^{pre} denote a specific bus queueing time, W denote passenger activity time and T^{post} be the merging time into the traffic. With this we can roughly write the dwell time matrix as

$$D = T^{pre} + W + T^{post} \quad (5.1)$$

T^{pre} is emphasized by the interactions between buses. This waiting time mainly depends on the length of the bus queue, the number of berths available and the passenger activity time of the previous buses. W is determined by passenger boarding and alighting activity. T^{post} refers to the traffic condition. A queue of bus vehicles is frequently observed at the stop shared by multiple bus lines, as is shown in Figure 5.1. The queue may consist of the buses purely from the same bus line, which is well known as the bus bunching problem. Sun and Schmöcker (2018) model W for the bus stop having two berths and allowing overtaking, which means that the bus that finishes loading/unloading first can instantly leave the bus berth. They consider the dynamics of the bus queue on a single line. In Figure 5.2, bus m is the bus of concern, a_m denotes the queue status at the arrival of bus m , and d_m denotes the

queue status at the departure of bus m . Sun and Schmöcker (2018) assume that passengers change boarding strategy according to the bus queue at the berth and overtaking policy. When the two berths are fully occupied such as the cases initiated by “2b” in Figure 5.2, the passengers may all turn to the first bus if overtaking is not allowed, and spontaneously split into two equal queues if overtaking is allowed so that the two buses leave the stop at the same time if the alighting process is negligible or finished before entering the berth. This interaction between passenger queue and bus vehicle queue is difficult to model and validate with real data. Schmöcker et al. (2016) consider the interaction between passenger queue and bus queue at a stop served by two lines. And they consider the bus queue at the passenger boarding/alighting area with two bus berths, so that there are only four patterns of bus queue: two buses from Line 1, two buses from Line 2, the bus of Line 1 after the bus of Line 2, the bus of Line 2 after the bus of Line 1. The patterns are exhaustively considered if the focus is narrowed within the passenger boarding/alighting area, as the resulting queue status will always fall into one of the four patterns when the first bus in the queue leaves and the third bus enters no matter which bus line they belong to. They further consider the choice set of the passengers given a simple bus network and analyze the effect of choice set and boarding strategy on W , although this method is not tested by complicated real bus network and shared stops. In contrast, Dai et al. (2019) model the bus queue within and outside the boarding/alighting area for real bus stops which are sometimes shared by four or five bus lines, but they does not consider the queue dynamics and passenger choice behavior.

W is widely studied and a brief review has been provided in Chapter 4. Linear or nonlinear regression is commonly applied to map the relationship between total passenger activity time and passenger boarding/alighting flows. Let p denote the index of method of payment, the total passenger activity time for a specific bus run at a given stop (omitting the bus index and stop index for simplicity) can be roughly written as in Eq. 5.2.

$$W = \begin{cases} \max\left(\sum_p t_p^a A_p, \sum_p t_p^b B_p\right) & \text{simultaneous case} \\ \sum_p t_p^a A_p + \sum_p t_p^b B_p & \text{sequential case} \end{cases} \quad (5.2)$$

In this chapter, we aim to investigate the interactions between buses in T^{pre} , the effect of in-vehicle crowding and different methods of payment on W . The time components and additional information

needed to understand the interactions and effects are discussed in Section 5.3. In Section 5.4, we discuss the onboard survey that is conducted to collect the data required to calibrate the dwell time models. In Section 5.5, the calibrated models and implications are discussed. In Section 5.6, the calibrated dwell time model is incorporated into the methodology provided in Chapter 4, in order to test the estimation performance on a busy and complex bus route. A summary for this chapter is in Section 5.7

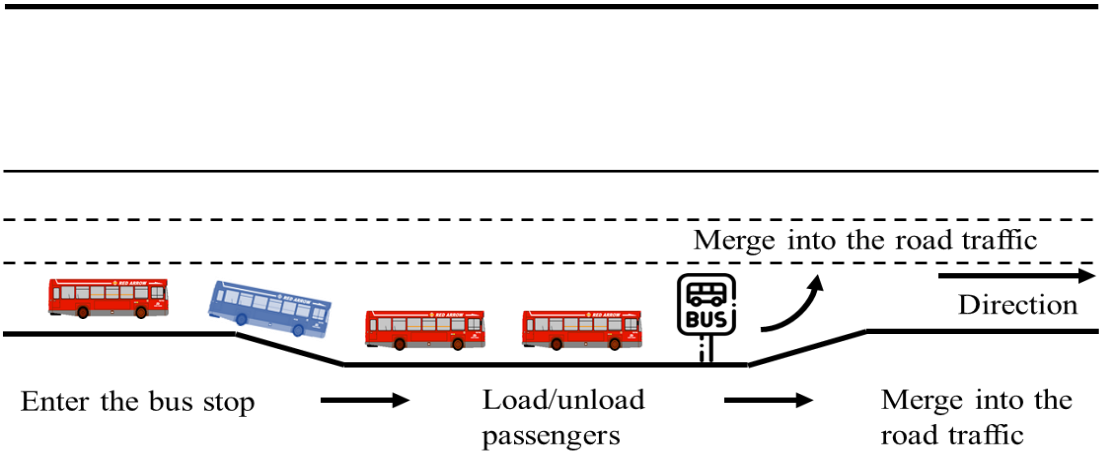


Figure 5.1 Bus dwelling process – Dai et al. (2019)

Case	Event	Illustration
1	a_m	
	d_m	
1→2f	a_m	
	d_m	
2b→1	a_m	
	d_m	

Case	Event	Illustration
2b→1→2f	a_m	
	a_{m+1}	
	d_m	
2b	a_m	
	d_m	
2b→3m→2f	a_m	
	a_{m+1}	
	d_m	
2b→3m	a_m	
	d_m	
3b	a_m	
	d_m	

Figure 5.2 Event sequences from the viewpoint of bus m - Sun and Schmöcker (2018)

5.3 Information needed to understand the bus dwell time

On the interactions between buses

According to the definition of the bus dwell time, the basic time components are arrival time and

departure time. For additional time components, continuous spatiotemporal coordinates (high recording frequency, less than 5 sec) are rarely available albeit ideal, the time stamps of stopping, accelerating to move, door opening and closing are very helpful for inferring the queue dynamics of the bus.

On the effect of in-vehicle crowding

Onboard number provides crucial clue on the effect of crowding. Onboard number at each stop of a bus route can be obtained if the streams of boarding and alighting data at each stop are available. We therefore need to collect the number of total boarding and alighting passenger for each bus stop, in addition to the time components. APC data is a reliable data source, and AFC data is also reliable if the percentage of smart card users is dominant. In addition, the crowding may result from the suitcases carried by the passengers on the bus route connecting city center and transportation hubs such as railway station and airport. This information can be observed at the stop or in the vehicle by survey or camera.

On the effect of methods of payments

A straightforward way is to record the number of the passengers using each method of payment. Onboard survey or in-vehicle camera is considered as the solution if a variety of methods are observed.

On the effect of other factors

The age of the passenger, the step at the entrance, and incidents such as wheelchair lifting and payment trouble may also affect the bus dwell time, which requires the data on the vehicle design details and passenger attributes. These effects are simplified in this research.

5.4 Onboard survey in Kyoto City

5.4.1 Overview of the survey

In order to collect the information discussed above, we conducted an onboard survey on three bus routes of different fare structures and passenger demand patterns in Kyoto City on Jan 26, 30 and 31, 2019. There are two fare structures observed Kyoto City. For many bus routes running in the downtown area, a flat fare structure is applied, whereas a distance-based structure for the bus routes in rural area or connecting the downtown and suburb. The bus fare is collected at the passenger's alighting. For the bus route of flat fare, no action is required at the boarding. For the bus route of

distance-based fare, a cash user has to take a number card based on which the actual fare is determined at the alighting, or to tap-in with a smart card. The effect of methods of payment is mainly observed during the alighting process. We select one bus route for each fare structure. In-vehicle crowding is frequently observed on the routes connecting famous tourist attractions, city center and the main railway station, we thus choose a bus route characterized by tourists as the third route to investigate. Table 5.1 summarizes the details.

Two investigators are assigned to each surveyed bus run, seated at the boarding and alighting door respectively. We assign investigators for successive three bus runs and ask them to stay in the vehicle for three round trips. Therefore we collect the data of nine bus runs for each direction each bus route. Table 5.2 and 5.3 list the data collected at the boarding and alighting door. In Kyoto City, passengers board at the back door and alight at the front door. For the back (boarding) door, data collection mainly focuses on the time components, as fare is collected at the alighting and it is almost impossible to observe the bus queue. For the front door, the data collection task is more challenging, passenger numbers distinguished by payment method as well as bus bunching status are required to record.

Table 5.1 Bus route investigated in the onboard survey

Bus route	Kyoto City Bus No. 3	Kyoto City Bus No. 12	Kyoto City Bus No. 29
Survey date	Jan 30 (Wednesday)	Jan 26 (Saturday)	Jan 31 (Thursday)
Fare structure	Flat fare	Flat fare	Distance-based fare
Demand pattern	Mostly residents	Mostly tourists	Mostly residents
Route configuration			

Table 5.2 Data collected at the boarding door

	Item	Description
1	Arrival time	hh:mm:ss
2	Boarding door open	hh:mm:ss
3	Boarding number (total)	
4	Boarding number (smart card)	
5	Boarding number (number card)	
6	Passenger with suitcase	
7	Boarding door close	hh:mm:ss
8	Departure time	hh:mm:ss
9	Road congestion	1: If the traffic congestion is observed from the previous stop to the stop of concern 0: otherwise
10	In vehicle crowding	1: seat available; 2: no seat 3: corridor occupied

Table 5.3 Data collected at the alighting door

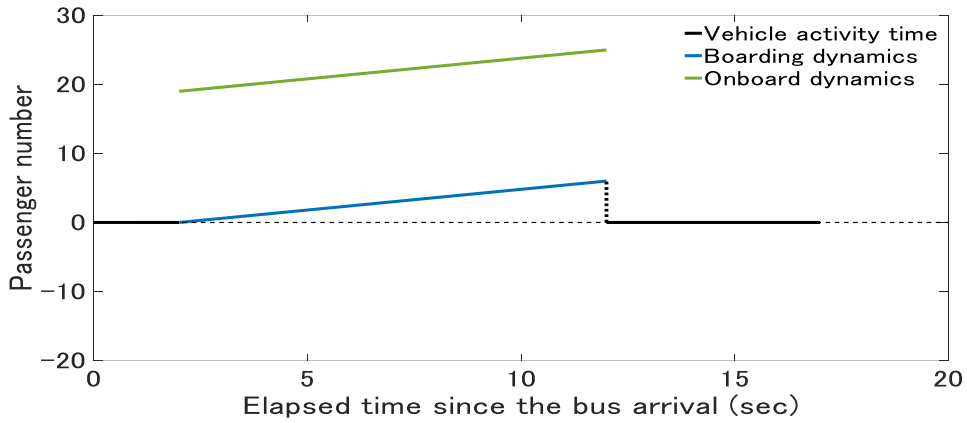
	Item	Description
1	Arrival time	hh:mm:ss
2	Alighting door open	hh:mm:ss
3	Alighting number (total)	
4	Alighting number (smart card)	
5	Alighting number (cash without in-vehicle exchange)	
6	Alighting number (cash with in-vehicle exchange)	
7	Alighting number (one-day ticket)	
8	Alighting number (one-day ticket activation)	
9	Alighting number (one-day ticket purchase)	
10	Passenger with suitcase	
11	Alighting door close	hh:mm:ss

12	Departure time	hh:mm:ss
13	Bunching between the buses of same line (single-line bunching)	1: bunching 0: otherwise
14	Bunching between the buses of different line (common-line bunching)	1: bunching 0: otherwise
15	Road congestion	1: If the traffic congestion is observed from the previous stop to the stop of concern 0: otherwise

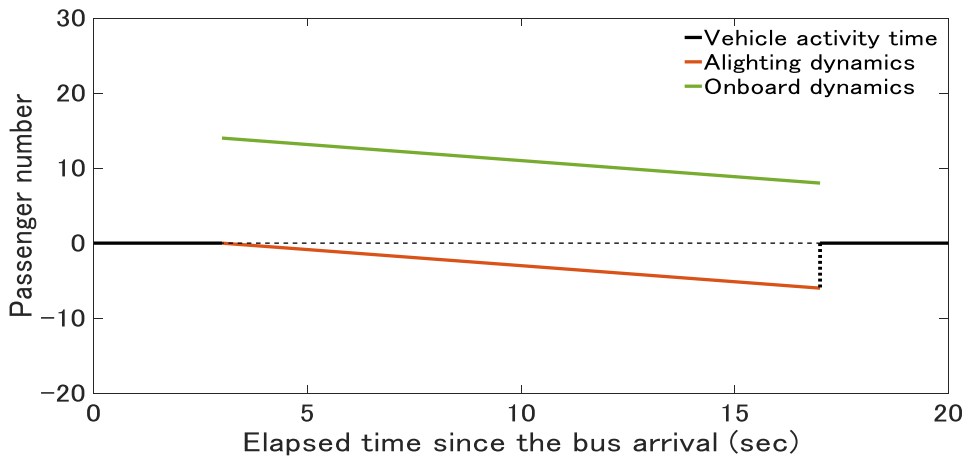
5.4.2 Descriptive analysis

We identify five distinctive sequences regarding passenger boarding and alighting flows, as are shown in Figure 5.3. Different from the time stamps left by smart card transactions, the real dynamics in the number of onboard, boarding and alighting passengers cannot be accurately captured from the survey data, the starting state and ending state of each number are thus simply connected to roughly demonstrate the changes. We focus on $T^{pre} + W$ which starts from the time point of a bus entering the bus area and the end of bus queue if it exists, and ends at the time point of the bus heading away from the bus stop. The time required to merge into the traffic is not collected. The bus stop in Kyoto City only has one berth, thus the second bus and further back ones in the bus queue can only unload the passengers while the first bus can load and unload the passengers. Figure 5.3 (a) and (b) show the boarding- and alighting-only patterns. In a normal dwelling, boarding and alighting proceed simultaneously as is shown in Figure 5.3(c), passenger activity time is hence determined by the longer one of boarding time and alighting time, referring to the simultaneous case in Eq. 5.2. Figure 5.3(d) illustrates a crowding case in which the alighting process begins ahead of boarding in order to release the space for the newly boarding passengers. The initial number of onboard passengers is 52, occupying 90% of the bus capacity. The bus capacity in Kyoto City is roughly 60, with on average 30 seats and a wide corridor in the front and a narrow corridor in the rear for standing passengers. We assume a constant boarding and alighting time per person to obtain the number of onboard passengers at the beginning of the boarding process, which is 40 in Figure 5.3(d). At last, a sequential case due to bunching is illustrated in Figure 5.3(e). The initial passenger load (23/60) is obviously below the crowding level though, a lagging of the boarding process can be observed. As the data regarding bunching (Item 13 and 14) at the alighting door refers to 1, it can be inferred that the bus unload the

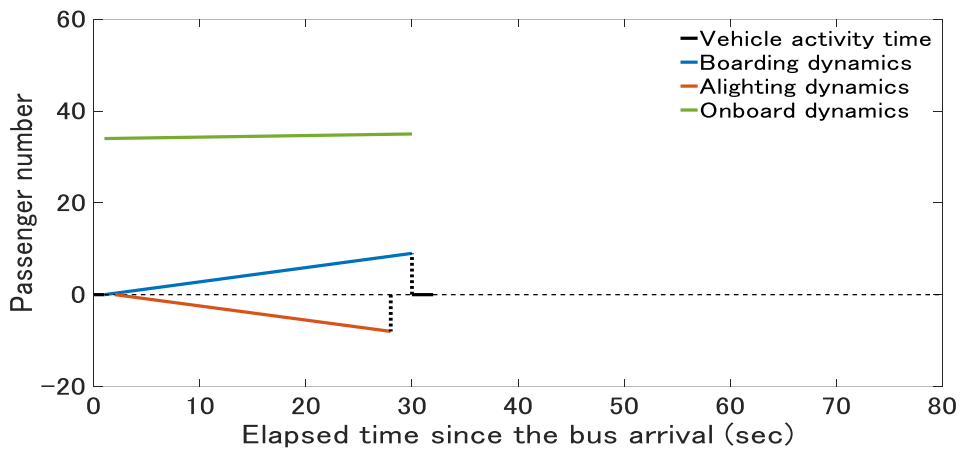
passengers at the bus queue and load passengers after the leading bus leaves the berth. The queueing time lasts for 14 sec, including the waiting time and the time for moving to the berth. These observed different sequences greatly help us to understand the complexity in the bus dwell process though, several critical pieces of information are not collected by the survey. Firstly, the investigator at the alighting door is seated behind the driver. Therefore it remains difficult to recognize the length of the bus queue and the route of the front bus due to the limited view horizon and the passing-through alighting flows. The records on single-line or common-line bunching are not reliable. The lack of data regarding the queue length at the arrival of the bus of concern makes it difficult to explain the variability in queueing time. Furthermore we find that Eq. (5.1) does not hold in Case (e), in that passengers alight during the queueing time, which increases the efficiency of bus dwelling under bunching circumstances.



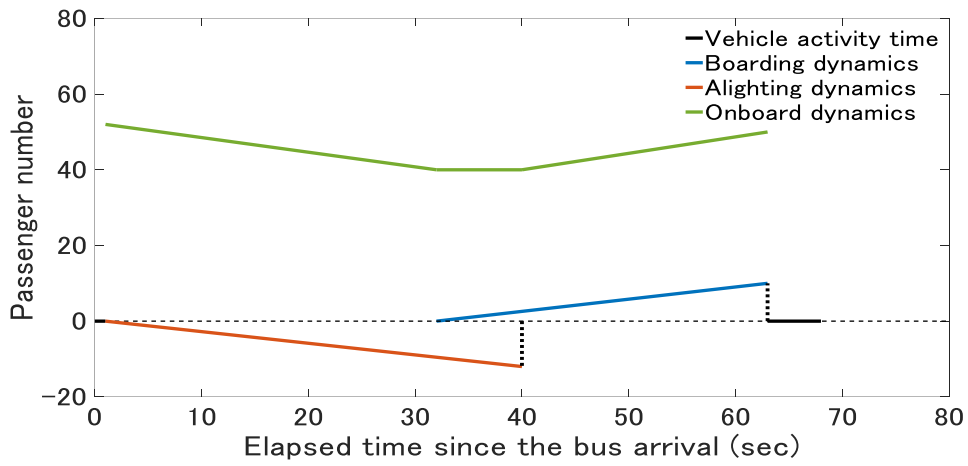
(a) Boarding only



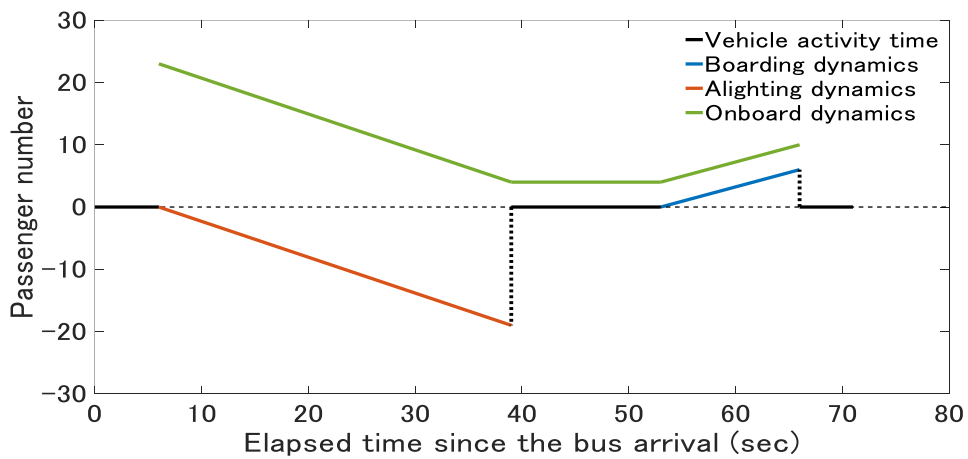
(b) Alighting only



(c) Simultaneous case



(d) Sequential case due to crowding



(e) Sequential case due to bunching

Figure 5.3 Different sequences of passenger boarding and alighting flows observed from onboard survey

5.5 Modeling the interaction between buses and the effect of in-vehicle crowding

We present the results based on the observations on Kyoto City Bus No. 12. The two directions of this route have different numbers of stops, 30 and 33 respectively. We therefore have 567 observations in total given nine bus runs. Three durations can be distinguished from the survey data, dwell time which is from bus arrival to departure, total boarding time which is from boarding door opening to closing, total alighting time which is from alighting door opening to closing. Firstly, we conduct the regression analysis by taking total alighting time and total boarding time as independent variables to investigate the relationship among bus dwell time, passenger activity time and vehicle activity time. Let W_A and W_B denote boarding and alighting time (scalar omitting the index of bus run and stop), ε be the vehicle activity time (dead time hereinafter), we have

$$D = \max(\beta_A W_A, \beta_B W_B) + \varepsilon \quad (5.3)$$

where β_A and β_B should significantly equal to 1.

We use Matlab to estimate the coefficients for this nonlinear regression, and Levenberg-Marquardt algorithm is employed. Levenberg-Marquardt algorithm solves nonlinear curve fitting by least squares, and it combines Gauss–Newton algorithm and Gradient descent. We set [1,1,1] as the initial value for the $[\beta_A, \beta_B, \varepsilon]$. 95% CI of the coefficients can be obtained by using the function “nlparci” in the statistical toolbox. This function returns the 95% CI for the nonlinear least squares parameter estimates with the residual and Jacobian matrix as the input. The standard deviation can be calculated with the 95% CI, then the standard error can be derived. Table 5.4 shows the regression results.

Table 5.4 Regression analysis on the bus dwell time, a basic model

	Estimate	SE	t-value	p-value
Dead time ε (sec)	7.9233	0.0280	283.2497	<0.001
β_A	1.0417	0.0014	733.0461	<0.001
β_B	1.0269	0.0019	552.7529	<0.001
$R^2 = 0.6807$				
Sample size = 567				

5.5.1 Modeling the effect of in-vehicle crowding

Although the general model fit is acceptable for this basic dwell time model, the dwell time models for cases (d) and (e) in Figure 5.3 should be distinguished. In 101 out of all 567 observations, neither boarding nor alighting passenger is observed. And we notice that many skipping cases also have dead time. The split for cases (a) to (e) in the other 466 samples is 159/129/125/28/25. Samples of case (d) are characterized by the crowding effect. We combine 125 normal simultaneous and 28 crowding simultaneous cases (153 cases in total) in the same model here. A dummy variable α_B is introduced to denote whether there are boarding passengers in the bus dwelling. N_{ON} is the number of onboard passengers at the arrival of the bus and β_{CT} is the crowding threshold. We assume that the boarding process will be delayed if the number of onboard passengers exceeds the threshold, and the crowding effect β_{CT} is in proportion of the exceeding amount. We set [1,1,10,1,1] as the initial values for $[\beta_A, \beta_B, \beta_{CT}, \beta_{CE}, \varepsilon]$.

$$D = \max(\beta_A W_A, \beta_B W_B + \alpha_B \max(N_{ON} - \beta_{CT}, 0) \beta_{CE}) + \varepsilon \quad (5.4)$$

The nonlinear regression problem is solved in the way as for Eq. (5.3). By taking the observations that the boarding process begins after the start of, but not the end of alighting process as the samples, we find that the boarding will be delayed if there are more than 29 passengers in the vehicle. And the lagging of boarding to alighting is positively correlated to the excessive number of passenger to the crowding threshold. Sun et al. (2015) define a critical occupancy to model the crowding effect, which is the ratio of onboard passenger to the bus capacity, and they observe the crowding effect when more than 60% of the capacity is occupied. Assuming that the capacity of the bus in Kyoto City is 60, the critical occupancy is then 50% in our case. It is not intuitively crowding. The boarding lagging tends to be generated under a relatively low critical occupancy in Kyoto, due to the fact that the standing passengers are likely to concentrate near the boarding door. The standing space in the rear part of the bus is narrow and requires the passengers to take a step.

Table 5.5 Regression analysis on the crowding effect in bus dwell time

	Estimate	SE	t-value	p-value
Dead time ε (sec)	5.9276	0.0730	81.1670	<0.001
β_A	1.0011	0.0023	443.7587	<0.001

β_B	1.0633	0.0040	267.7098	<0.001
Crowding effect β_{CE} (sec/pax)	0.1957	0.0077	25.4552	<0.001
Crowding threshold β_{CT} (pax)	29.1242	0.6751	43.1394	<0.001
$R^2 = 0.9096$				
Sample size = 153				

5.5.2 Modeling the interaction between buses

We combine 25 sequential cases, 159 boarding only cases and 129 alighting only cases (313 cases in total) in the same model here. α_{Bun} is a dummy variable to distinguish the bunching events. ε_{Bun} is the bunching effect, namely the dead time due to bunching. Considering that the boarding is likely to be delayed after the bunching effect, we have

$$D = \beta_A W_A + \alpha_{Bun} \varepsilon_{Bun} + \beta_B W_B + \alpha_B \max(N_{ON} - \beta_{CT}, 0) \beta_{CE} + \varepsilon \quad (5.5)$$

Table 5.6 reports the regression results which suggests that the bunching effect is 8.73 sec and the crowding threshold is close to the results in the simultaneous case. And in all the models, β_A and β_B are significant and close to unit value. In the next subsections, the regression analysis is conducted on W_A and W_B .

Table 5.6 Regression analysis on the effect of bunching and crowding in bus dwell time

	Estimate	SE	t-value	p-value
Dead time ε (sec)	5.9080	0.7816	93.9966	<0.001
β_A	1.0371	0.3264	344.3877	<0.001
β_B	1.0460	0.7716	332.1260	<0.001
Bunching effect ε_{Bun} (sec)	8.7276	1.2723	90.3097	<0.001
Crowding effect β_{CE} (sec/pax)	0.1190	0.5230	11.3669	<0.001
Crowding threshold β_{CT} (pax)	29.9025	0.3000	24.4523	<0.001
$R^2 = 0.6909$				
Sample size = 313				

5.5.3 Regression analysis on the alighting time

Here we use the data collected at the alighting door of Kyoto City Bus No.12 to investigate the effect of payment methods, as one-day ticket users who are mainly tourists are frequently observed on this route. The dataset includes the 307 cases when at least one passenger alight. As is shown in Table 5.7, there are various methods of payment observed on this route. It takes 2.12 sec for a smart card user to tap-out when alighting, which is consistent with the findings in other studies (1.58 - 2.37 sec/pax by Tirachini, 2013; 1.69 sec/pax by Sun et al., 2015). The buses in Japan are usually equipped with a cash-to-coin exchange machine, and the passengers may exchange before alighting or more commonly in the queue of alighting. The boarding time for a cash user is 4.10 sec/pax and 3.14 sec/pax more if exchange is needed. As is mentioned, this bus route is popular for the tourists, as it connects the city center and several of the most famous tourist attractions in Kyoto such as Kinkaku Temple (Golden Pavilion) and Gion. For the tourists, on-day ticket is a convenient and saving-money way. It requires to activate the ticket at the first time of using the ticket following the driver's instruction, which costs 4.24 sec/pax. Since the second time, the alighting time is reduced to 2.28 sec/pax which is close to the smart card user because the user only has to show the time-stamped ticket to the driver. Many places in the city are selling the one-day ticket, though, the passengers can buy it directly from the driver and meanwhile activate it when alighting the bus, the whole process costs 12.16 sec/pax which is close to the summation of time for exchange and activation. Other methods refer to the passengers using commuter/student pass, or the payment is not identified by the investigators. We observe 1208 passengers, the percentage of each payment is listed in Table 5.7. The weighted alighting time per person is 2.14 sec. Although the overall model fit is positive, the time per person for other methods which is also the largest proportion appears underestimated. This might result from those passengers alighting the bus when other passengers are making the payment, since they keep the segment pass and do not need to pay, also many of them might be wrongly classified due to manual mistakes. We conduct another simple regression to obtain the average alighting time per person as in Table 5.8.

Table 5.7 Regression analysis on alighting time W_A

	Estimate	SE	t-value	p-value	Percentage
Dead time ε_A (sec)	8.8019	0.7816	11.2613	<0.001	
Smart card (sec/pax)	2.1174	0.3264	6.4862	<0.001	29.14%
Cash	4.0984	0.7716	5.3117	<0.001	8.48%
Cash with onboard	7.2366	1.2723	5.6878	<0.001	2.55%

exchange					
One-day ticket (activate)	4.2418	0.5230	8.1109	<0.001	9.05%
One-day ticket (show)	2.2837	0.3000	7.6127	<0.001	19.75%
One-day ticket (purchase)	12.1638	2.4477	4.9695	<0.001	0.58%
Other methods	0.2810	0.4186	0.6713	0.5026	30.45%
Adjusted $R^2 = 0.7128$					
Sample size = 307					

Table 5.8 Regression analysis on alighting time W_A (without distinguishing methods of payment)

	Estimate	SE	t-value	p-value
Dead time ε_A (sec)	5.9978	0.8545	7.0189	<0.001
Alighting (sec/pax)	2.8648	0.1403	20.4169	<0.001
Adjusted $R^2 = 0.5761$				
Sample size = 307				

5.5.4 Regression analysis on boarding time

Here we use the data collected at the boarding door of Kyoto City Bus No.12. Model fitting is fairly unfavorable. The dataset includes the 337 cases when at least one passenger boards. It can be observed in reality that the boarding door sometimes remains open after the bus finished loading the passengers at the stop and is closed together with the alighting door, which explains why the intercept is longer than that in the alighting time model. As no action is required at the boarding of the bus, the time interval between two successive boarding passengers varies significantly from 0 (two passengers board at the same time as the boarding door is wide enough) to more than 10 seconds due to carrying large suitcases. The dependency on the alighting time and the variation in individual boarding account for the low goodness of fit of the boarding time model.

Table 5.9 Regression analysis on boarding time W_B

	Estimate	SE	t-value	p-value
Dead time ε_B (sec)	9.5948	1.0083	9.5155	<0.001
Boarding (sec/pax)	1.3316	0.2026	6.5715	<0.001
Adjusted $R^2 = 0.1142$				
Sample size = 337				

5.6 Application of the complex dwell time model to OD estimation framework

5.6.1 *Data detail level required and techniques for data processing*

We successfully capture the bunching and crowding effect via an onboard survey in the previous section, and in this section we introduce the steps to incorporate the complex dwell time models into the OD estimation framework proposed in Chapter 4.

For the bus line always faced with bunching and crowding, it is essential to identify the ineffective part in the dwell time in which there is no passenger flow. Given two buses bunching at the stop, from the perspective of the back bus, the forward headway is short while the dwell time is long as it has to wait for the front bus. Not considering the bunching effect may significantly overestimate the passenger arrival rates given the short headway and long dwell time. Besides, a simultaneous form of the dwell time model is applied to a normal case though, a sequential form should be applied to a bunching case after excluding the dead time due to bunching. Furthermore, the passengers may always board the leading bus in the single-line bunching case if overtaking is not allowed and there is space remaining in the first bus. In order to identify the bunching between the buses from different lines, the AVL data of a single bus route appears not enough. For the bus stop where multiple lines are intercepted, the joint headway should be used for the bunching judgement.

It is difficult to precisely exclude the dead time due to bunching. If there is a bus queue, we suppose that the bus initial arrival time is the time point of the bus entering the bus queue, and that the passengers can alight after the dead time required for door opening. The bus closes the door when it finishes unloading passengers, and cannot move to the berth to load passengers until its leading bus leaves the bus berth. The bus has to move two times if it queues after two buses. The time used for waiting and moving depends on the queue length. The bus departs after the passenger boarding time at the berth and door closing time. The entire dwell time from the initial arrival to the final departure should be separated by several time points, in order to obtain the actual time for passenger boarding and alighting:

- (1) Initial arrival time, or alighting door opening
- (2) End of the alighting process, or alighting door closing
- (3) Arrival time at the bus berth, or boarding door opening

(4) Final departure time, or boarding door closing

Time interval between points (1) and (2) is regarded as the alighting time, and that between (3) and (4) is the boarding time. The accuracy of the split strongly depends on the detail level of the dataset. Refer to Table 2.1, detail level C records the arrival and departure time of the bus at each stop, which are points (1) and (4). Points (2) and (3) have to be inferred. It is straightforward to obtain the end of the alighting time and the start of boarding if AFC data is available, and it is also feasible to infer these time points if the AVL data is of high detail level, such as level E (every 5 seconds or more frequent), or recoding every door opening and closing or containing speed information. For the AVL data of low detail level, such as every 10 or more seconds, it is difficult to determine all four time points, however we may narrow the time window of the time points if the data of other buses involved in the same queue are collectively considered.

5.6.2 *A case study in Kyoto City*

Refer to Chapter 4, passenger activity time and headway obtained from bus AVL data are the main model inputs. Besides, boarding and alighting time per passenger are also assumed for the model. In the case study in Chapter 4, 1.5 sec/pax is assumed for boarding and 2.5 sec/pax for alighting. In this section we prepare data regarding passenger activity and headway with the rich time points provided by the survey data. And we apply the fitted complex model in the last section. Kyoto City is a good candidate to investigate the effect of dwell time complexity on passenger flow estimation, considering that frequent bus bunching between the buses either from the same line or different lines occurs. With the help of bunching information collected by the survey, we can apply simultaneous form or sequential form to the bus dwell time properly. Another reason of using the survey data is that the ground truth of passenger flows are available. Although plentiful bus AVL data is available in Kyoto City though, we do not have the access to APC and AFC data. We also note the shortcoming of the survey dataset that it records the trips of nine bus runs for each bus route however from 8am to 16pm which involves several heterogeneous time-of-day intervals. The expected passenger flows might be biased by these sampled trips. As is discussed in Chapter 4, the bus trips in a homogeneous observational period are supposed to be used to estimate the mean arrival rate for that time period, considering the time-dependency of passenger flows. Finally, we simplify the crowding effect to reduce the computational complexity. For comparison purpose, two dwell models are applied in the demand estimation of this chapter. The advanced one is as in Eq. (5.6), which uses weighted alighting

time per person and considers bus bunching. The simple one is the above part of Eq. (5.6), and it uses the alighting time per person not weighted by methods of payment.

$$D = T^{Pre} + W = \begin{cases} \max(t^b B + \varepsilon_B, t^a A + \varepsilon_A) + \varepsilon_1 & \text{bunching does not occur} \\ t^a A + \varepsilon_A + \varepsilon_{Bun} + t^b B + \varepsilon_B + \varepsilon_2 & \text{bunching occurs} \end{cases} \quad (5.6)$$

Refer to Table 5.6, 5.8 and 5.9, we have the coefficients as shown in Table 5.10 for simple dwell time model and advanced dwell time model.

Table 5.10 The simple and advanced dwell time models applied in demand estimation

	Simple dwell time model	Advanced dwell time model
$t^b =$	1.33 sec/pax	1.33 sec/pax
$t^a =$	2.86 sec/pax	2.14 sec/pax
$\varepsilon_1 =$	8 sec	8 sec
$\varepsilon_2 =$	N/A	6 sec
$\varepsilon_A =$	6 sec	9 sec
$\varepsilon_B =$	10 sec	10 sec
$\varepsilon_{Bun} =$	N/A	9 sec

Four scenarios are distinguished in this case study: AB, BT, T and TB. The settings for AB, BT and T are identical to those in Chapter 4. The simple dwell time model is used in BT. The difference between T and TB is that TB incorporates the advanced dwell time model and T uses the simple one. We remove AT due to the confirmed underperformance of AT to BT. Here we only report the estimation performance on passenger boarding, alighting and onboard numbers, due to the lack of ground truth on OD flows.

The data of one direction of Route 12 (30 bus stops) is used. In the 9-bus-run dataset, single-line bunching does not occur and common line bunching occur 25 times in 270 bus dwelling events. Figure 5.4 to 5.6 show the estimation performance on boarding, alighting onboard number of passengers in four scenarios. Due to the diversity of time-of-day intervals and the lack of sample trips, significant variation can be observed in the ground truth. Although the mean passenger flows in both ground truth and estimation result should not be interpreted as the true mean flows of the day, we still can draw

several important conclusions. The emphasis is put on the effect of dwell time models in this case study, which refers to the comparison between T and TB. As is illustrated in Table 5.12, TB outperforms T slightly in terms of boarding flows, and significantly in terms of passenger loads. TB even underperforms T regarding alighting flows. This indicates that the special treatment on bunching improves the estimation on passenger boarding flows and passenger loads, however that on payment methods may not improve the estimation on passenger alighting flows. Advanced dwell time model can greatly increase the accuracy of passenger load and crowding level estimation though, the improvement for boarding flows is limited. This may results from that bunching is not severe and not a systematic problem to the selected bus route. A bunching rate of 25/270 is probably not high enough to fundamentally affect the estimation. The data of a bus route which is more characterized by bus bunching is needed to verify this conclusion. However, the impact of bunching on passenger flow reconstruction should be significant. Passenger flows reconstruction is to match the passenger flows for each bus run. This is an interesting further step of this research.

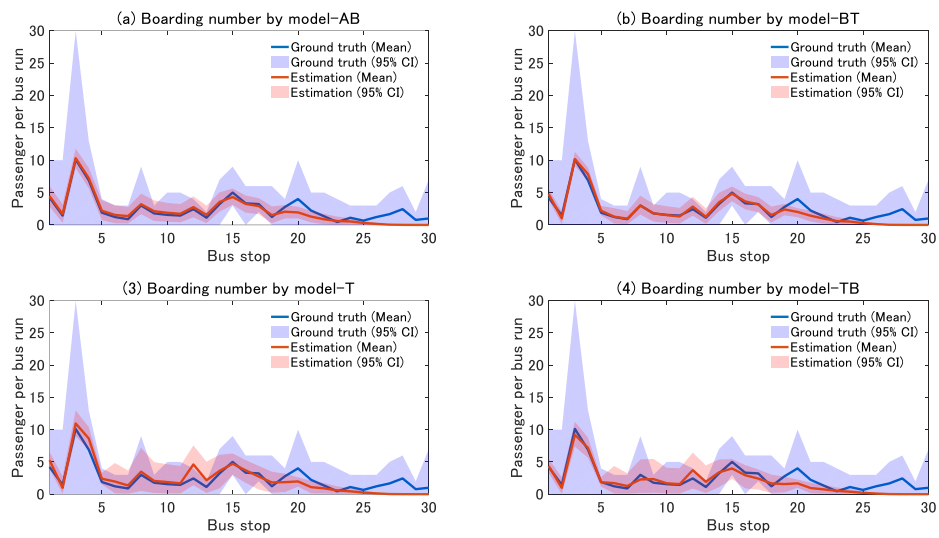
Scenario AB produces reliable estimation for the bus route in Shizuoka City, as is shown in Table 5.12. The results for the bus route in Kyoto City are unfavorable. Again we note that the time-dependency is not considered. Taking stop 3 Kinkaku-ji mae as an example, the boarding flows vary from [1,0,1] for the first 3 runs in the morning to [30,13,13] for the last three runs in the survey in the afternoon, as many tourists finish their visits to this famous temple and take a back trip to city center in the afternoon. Therefore, the estimation performance can be greatly improved if more data is available.

Table 5.11 The prior knowledge on the stops of Kyoto City Bus No. 12

		Prior (AB, BT)	Prior* (T, TB)	Stop characteristics
$p_i^g, i =$	1	$U[0,10]$	$U[2,3]$	Ritsumeikan University
	3	$U[0,10]$	$U[6,7]$	Kinkaku-ji mae
	4	$U[0,10]$	$U[5,6]$	Kinkaku-ji michi
	20	$U[0,10]$	$U[8,9]$	Nijo Castle
	21-30	$U[0,10]$	$U[0,10]$	
$p_i^a, i =$	1-30	$U[0,10]$	$U[0,10]$	

Table 5.12 RMSE and MAPE of the estimation results for each scenario

Model	Boarding flows		Alighting flows		Passenger loads	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
AB (Shizuoka Route A)	0.1964	7.60%	0.1421	7.36%	1.0964	4.36%
AB (Kyoto No. 12)	0.8114	24.32%	0.5814	13.88%	2.7883	8.10%
BT	0.7924	21.57%	0.8898	27.28%	2.9943	8.42%
T	1.0296	34.14%	0.9564	30.47%	5.9337	17.45%
TB	0.9626	30.90%	1.1725	33.70%	3.9830	9.55%

**Figure 5.4** Estimated passenger boarding flows

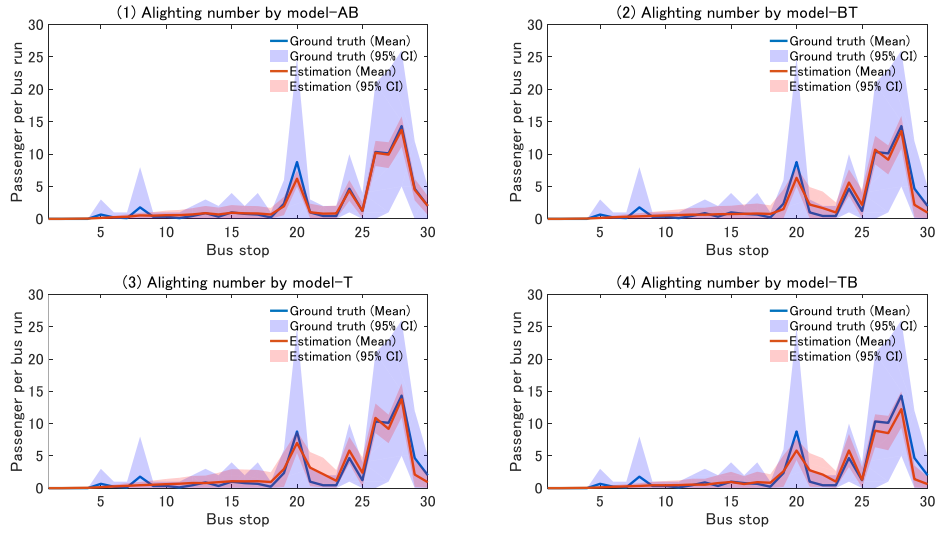


Figure 5.5 Estimated passenger alighting flows

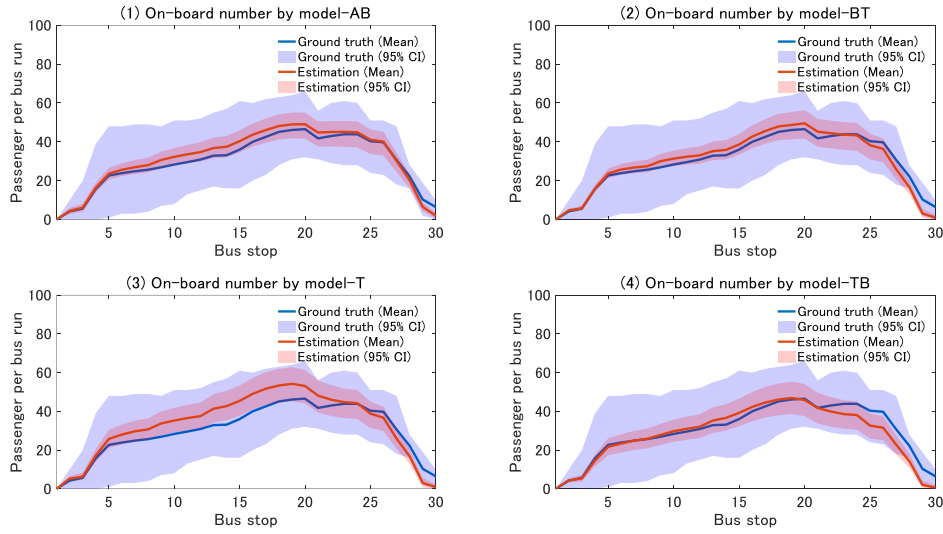


Figure 5.6 Estimated passenger loads

5.7 Summary

In this chapter, a series of dwell time models are provided to investigate the effect of crowding and bunching on the bus dwell time. We conduct a survey in Kyoto City to collect the data and verify the models. Five distinctive patterns are observed from the survey data, in particular the crowding simultaneous case and bunching sequential case. Furthermore, the fitted dwell time models are used as input to the estimation framework on passenger flows which is developed in Chapter 4. We find

that an advanced dwell time model can improve the estimation on boarding flows and significantly improve the estimation on passenger loads. This limited improvement on boarding flows may result from the bunching problem on the selected bus route is not severe enough to impose impacts on the estimation.

Besides, several practical findings should be noted. Firstly, the shortcomings in the existing ticketing policy is pointed out. Alighting time per person by cash exchange and one-day-pass purchase are 7.24 sec and 12.16 sec respectively. These two types of payment greatly prolong total alighting time and delays bus departure time. The exchange machine is located beside the fare collection box at the alighting door and the pass purchase is based on the communication between passenger and driver, both of which blocks the way of other alighting passengers. New ticketing policies that make it more convenient for the passengers to conduct cash exchange and pass purchase before the bus arrives their destinations are expected to mitigate the delays. Secondly, the negative effect of bus bunching is quantified. The process of bus queueing to the loading/unloading area due to bus bunching and lane congestion last for 8.7 sec on average. As bunching tends to keep delaying the bus for successive stops, and the delay effect spreads to following buses, the accumulated delay can be fairly significant. We hence call attention on the countermeasures against bus bunching, and the bunching prediction tool proposed in this thesis can make contributions.

Chapter 6 **Conclusions**

6.1 Summary of research

The main objective of this research is to illustrate the capability and potential of bus AVL data terms of assisting the decision making in transit planning and operation. The emergence of transit data has significantly provided the possibility for them to understand the status quo of the service, the insufficient part in particular; also to encounter the service inadequacy in a data-drive way. However, AVL data is not the only option for the operators who are trying to establish data-driven planning and operation system. The existing literature has already presented the advantages of AVL data in obtaining time components in the bus trips, and also the disadvantages of capturing the information regarding demand and passenger flows. Multiple data sources such as APC and AFC data on the other hand, contain plentiful demand information and the key time components of a bus trips such as arrival and departure time. As a result, advanced methodologies on data-driven planning and operation require APC or AFC data, or the collaboration of multiple data sources, which imposes the expression that the contribution that AVL data can make is limited. This may enlarge the gap in the service quality between “data-rich” operators, that is, those who have access to multiple data sources, and “data-poor” operators who may only have AVL data. This research therefore considers the problem from the perspective of the “data-poor” operators (there are “data-poorer” operators that have no access to data though, they are beyond the scope of this research).

Chapter 2 compares the three main-stream transit data: AVL, APC and AFC, comprehensively explain the capability of and existing literature regarding these datasets, as well as the possible contribution of collaboration among them. The data-driven applications planning, service analysis, prediction and control in real-time and so forth. We draw the conclusion that bus AVL data can be used in predicting arrival time, departure time, travel time, headway, etc. for bus trips, in measuring the punctuality and headway evenness, and also in identifying distinctive time-of-day operating intervals. The main inferiority of AVL data over the other two datasets is that it barely convey demand information. Precisely speaking, the passenger-related information concealed in bus AVL data is way from being fully undermined. We thus consider this as a crucial research direction and also a main objective of this research.

Chapter 3 provides a relatively conventional research direction of taking the advantage of AVL data.

A logistic regression model is developed to predict bus bunching in multiple-step predictions. The effective prediction can be as long as 15 stops. The model input is the time components extracted from AVL data such as dwell time and headway at an upstream stop distanced from the stop of concern, and the model output is the probability of bunching to occur at the stop of concern. The probabilistic prediction results are more robust and reliable than the exact values produced by headway-based approaches.

Chapter 4 is the core part of this research. It undermines the potential of AVL data in inferring passenger demand. A novel methodology, to our knowledge also the first attempt, is proposed to estimate passenger OD flows using AVL data. We model the demand with passenger arrival rate per minute at OD-pair level. With the headway and dwell time obtained from AVL data, we can relate the passenger flows to the bus dwell time. A Bayesian inference framework is applied for this underspecified problem. And a Hamiltonian MCMC algorithm is employed to draw the distribution for the unknown parameters. The inferred mean and confidence interval are reliable estimates for operators to figure demand-driven bus route design and frequency.

Chapter 5 enhances the approach presented in Chapter 4 in that the complexity in bus dwell time is now considered. It considers the entire bus dwell process at a stop, which consists of pre-loading/unloading stage (bus queueing to enter the bus berth), loading/unloading stage (at the bus berth), post-loading/unloading stage (bus leaving the berth to merge into the traffic flow). We find in a Kyoto City survey that sometimes the separation of pre-loading/unloading stage and loading/unloading stage is implicit, and the unloading (passenger alighting) activity may proceed when the bus is waiting in the bus queue. By applying a fitted bus dwell time model considering bus queueing which is also known as bus bunching to the estimation framework in Chapter 5, the estimation performance in boarding flows and passenger loads is improved, and the latter one is significant. It is worth of further investigation of whether it significantly matters when it comes to the passenger flow reconstruction problem whose goal is to match the exact number for each bus run. Chapter 5 also quantifies the effect of various payment methods and bus bunching on service delays.

6.2 Contribution to existing knowledge

Firstly, the literature on the characteristics of these widely used transit data sources, on the data-driven methodologies serving for transit planning and operation based on these datasets, in particular bus

AVL data, are systematically reviewed. A full picture of the data sources, of the advantages and disadvantages of bus AVL data, of the existing literature is an initial and important contribution of this research.

Secondly, the probabilistic prediction tool developed in this research provides the bus operators tradeoff options, which enables them to project flexible bus control in real time. It is advantageous in predicting bus bunching events comparing the traditional methods that transform the predicted exact value of headway to binary result. It provides fresh insight in the knowledge domain of prediction methods for bus transit. Also the reliability in long-term prediction (prediction horizon is more than 10 stops) is noteworthy.

Thirdly, the estimation framework on passenger OD flows using AVL data may substantially broaden the minds of bus operators and the contribution to the existing knowledge is both methodological and practical. We model the demand with OD-pair passenger arrival rate. The estimated rates can be easily transformed to stop-level arrival rates and alighting probabilities, which are the input of many advanced bus control models. The model inputs are usually calibrated from APC, AFC or onboard survey data. Now, the demand parameters and time component parameters such as mean inter-stop travel time, mean headway, etc. can be obtained from a single AVL dataset. We note and illustrate that we are not suggesting that a single dataset outperforms the collaboration of multiple datasets. Instead, we are pointing out that the situation confronting the “data-poor” operators should be seriously considered.

Fourthly, this research models the complex bus dwell process with a consideration of in-vehicle crowding, bus bunching and lane congestion. It successfully spotlights the shortcomings of existing ticketing policies in Kyoto and quantifies the negative effect of bus bunching on service delays. New policies regarding onboard cash exchange and pass purchase, as well as predictive control strategies towards bus bunching have great potential in improving the existing service.

6.3 Future research directions

For the probabilistic prediction tool developed in Chapter 3, some extensions are supposed to strengthen the predictive power of the model. The model itself has space for improvement by including traffic signals and passenger demand into the set of independent variables. They are not incorporated

in this research due to the lack of data. Furthermore, the study could be extended to simultaneously predict bunching for several lines, in which case common line effects such as the interaction between buses of different lines at a common stop need to be considered. It is challenging to model the bus bunching for the stops at which multiple lines are intersected and the route segment served by multiple lines.

In Chapter 4, we discuss the difference between mean OD matrix estimation and matrix reconstruction. The estimated mean passenger flows and passenger rates are useful in bus route planning in terms of route design and timetable setting. The matrix reconstruction is then put the focus on reproducing the passenger flows and load profile for each bus run using AVL data and estimated passenger rates, which is a two-stage estimation. It is expected to provide the estimates on how worse the real situation is, the fluctuation of the passenger flows from the expectation. Instead of the distribution of expected (mean) passenger flows, the distribution of actual passenger flows can be obtained.

In Chapter 5, we conclude that the consideration of bunching in an advanced dwell time model does not significantly improve the estimation on passenger flows. We suppose this impact to be significant in reproducing the passenger flows. The gap in the passenger boarding flows at the bunching stops and passenger loads over the bunching segment between two successive bunched buses should be obvious, and this inequality is not captured when we estimate the mean value.

Another noteworthy methodological extension of this estimation framework is to make the estimation fully independent of APC data. In Chapter 4, we assume stop generation power and attraction power to model the passenger OD-pair arrival rates and reduce the number of unknown parameters. We infer the prior knowledge for these unknown parameters in terms of the lower and upper bounds for the stop importance from APC data. This passenger count data can be alternative obtained from a small sample survey though, the method to refine stop importance is worth exploring. The data sources that are conventionally used to calibrate such information, such as population data, are not easy to collect at a bus stop base. We thus provide some candidate data here that have the potential to explain the generation and attraction power of the stop, and can be collected at a low cost. Number and property of the POIs (Points of Interest) within a certain distance from the bus stop are considered powerful in explaining the importance of stops, and they can be freely obtained from Open Street Map or other map databases. In addition, the topology of the bus transit network and the connectivity of the bus stop

are strong supplements to obtain importance weights. Considering the massive types of POIs, the complex topology of transit networks and a need to infer demand influence radius of stops, deep learning approaches such as graph convolutional network can take the place of gravity model we incorporate in this framework to determine the weight of POIs and the distance cost function. Koca (2020) estimated the OD matrix of taxi trip data for New York using “OpenStreetMap” data in combination with deep learning approaches. The combination of AVL data and map data, and a hybrid approach of deep learning and Bayesian inference may enhance the transferability of the estimation framework for “data-poorer” operators.

Last but not least, the demand is modeled with OD-pair arrival rates which can be used to simulate the passenger demand in the bus propagation/evolution models. In bus control models, the controlling performance is usually evaluated by the future states of the bus system, and the future states can be derived by the bus propagation model. Stop arrival rates and alighting probabilities are often assumed for these models and calibrated from APC data in the case study though, the assumption of OD-pair arrival rates is closer to the real passenger behavior. Few studies assume OD-pair rates as it can be only calibrated from AFC data or OD survey. Now we make it feasible to obtain OD-pair rates from bus AVL data, significantly reduce the calibration difficulty. We expect more studies can consider passenger demand at OD level based on the framework we propose and develop more effective strategies to improve bus service.

Reference

- Andersson, P. Å., Hermansson, Å., Tengvald, E., & Scalia-Tomba, G. P. (1979). Analysis and simulation of an urban bus route. *Transportation Research Part A: General*, 13(6), 439-466.
- Andres, M., & Nair, R. (2017) A predictive-control framework to address bus bunching. *Transportation Research Part B*, 104, 123-148.
- Arriagada J., Gschwender, A., Munizaga, M., & Trepanier, M. (2019). Modeling bus bunching using massive location and fare collection data. *Journal of Intelligent Transportation Systems*, 23(4), 332-344.
- Barry, J. J., Newhouser, R., Rahbee, A., & Sayeda, S. (2002). Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record*, 1817, 183-187.
- Bartholdi, J. J., & Eisenstein, D.D. (2012). A self-coordinating bus route to resist bus bunching. *Transportation Research, Part B: Methodological*, 46(4), 481-491.
- Bell, M. G. (1991). The estimation of origin-destination matrices by constrained generalised least squares. *Transportation Research Part B: Methodological*, 25(1), 13-22.
- Ben-Akiva, M., Macke, P. P., & Hsu, P. S. (1985). Alternative methods to estimate route-level trip tables and expand on-board surveys. *Transportation Research Record*, 1037, 1-11.
- Berrebi, S. J., Watkins, K. E., & Laval, J. A. (2015). A real-time bus dispatching policy to minimize passenger wait on a high frequency route. *Transportation Research Part B*, 81, 377-389.
- Berrebi, S. J., Hans, E., Chiabaut, N., Laval, J. A., Leclercq, L., & Watkins, K. E. (2018). Comparing bus holding methods with and without real-time predictions. *Transportation Research Part C*, 87, 197-211.
- Betancourt, M., & Girolami, M. (2013). Hamiltonian Monte Carlo for hierarchical models. *arXiv preprint arXiv:1312.0906*.
- Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.
- Bian, B., Zhu, N., Ling, S., & Ma, S. (2015). Bus service time estimation model for a curbside bus stop. *Transportation Research Part C: Emerging Technologies*, 57, 103-121.
- Bie, Y., Gong, X., & Liu, Z. (2015). Time of day intervals partition for bus schedule using GPS data. *Transportation Research Part C: Emerging Technologies*, 60, 443-456.
- Bie, Y., Xiong, X., Yan, Y., & Qu, X. (2020). Dynamic headway control for high-frequency bus line

- based on speed guidance and intersection signal adjustment. *Computer-Aided Civil and Infrastructure Engineering*, 35(1), 4-25.
- Bookbinder, J. H., & Desilets, A. (1992). Transfer optimization in a transit network. *Transportation science*, 26(2), 106-118.
- Camus, R., Longo, G., & Macorini, C. (2005). Estimation of transit reliability level-of-service based on automatic vehicle location data. *Transportation research record*, 1927(1), 277-286.
- Cascetta, E., & Nguyen, S. (1988). A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research Part B: Methodological*, 22(6), 437-455.
- Ceder, A., & Wilson, N. H. (1986). Bus network design. *Transportation Research Part B: Methodological*, 20(4), 331-344.
- Cham, L. C. (2006). Understanding bus service reliability: a practical framework using AVL/APC data (Master's thesis, Massachusetts Institute of Technology).
- Chan, J. (2007). Rail transit OD matrix estimation and journey time reliability metrics using automated fare data (Master's thesis, Massachusetts Institute of Technology).
- Chang, H., Park, D., Lee, S., Lee, H., & Baek, S. (2010). Dynamic multi-interval bus travel time prediction using bus transit data. *Transportmetrica*, 6(1), 19-38.
- Chen, M., Liu, X., Xia, J., & Chien, S. I. (2004). A dynamic bus arrival time prediction model based on APC data. *Computer-Aided Civil and Infrastructure Engineering*, 19(5), 364-376.
- Chen, X., Yu, L., Zhang, Y., & Guo, J. (2009). Analyzing urban bus service reliability at the stop, route, and network levels. *Transportation research part A: policy and practice*, 43(8), 722-734.
- Chen, X., Hellenga, B., Chang, C., & Fu, L. (2015). Optimization of headways with stop-skipping control: A case study of bus rapid transit system. *Journal of Advanced Transportation*, 49(3), 385-401.
- Chen, X., He, Z., & Sun, L. (2019). A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation. *Transportation research part C: Emerging technologies*, 98, 73-84.
- Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks* 17(1), 113-126.
- Chien, S. I. J., Ding, Y., & Wei, C. (2002). Dynamic bus arrival time prediction with artificial neural networks. *Journal of Transportation Engineering*, 128(5), 429-438.
- Coffey, C., Pozdnoukhov, A., & Calabrese, F. (2011, November). Time of arrival predictability horizons for public bus routes. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Computational Transportation Science* (pp. 1-5). ACM.

- Cortés, C. E., Jara-Díaz, S., & Tirachini, A. (2011). Integrating short turning and deadheading in the optimization of transit services. *Transportation Research Part A: Policy and Practice*, 45(5), 419–434.
- Cosslett, S. R. (1981). Maximum Likelihood Estimator for Choice-Based Samples. *Econometrica*, 49(5), 1289-1316.
- Daganzo, C.F. (2009). A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Research Part B*, 43(10), 913-921.
- Daganzo, C.F., & Pilachowski, J. (2011). Reducing bunching with bus-to-bus cooperation. *Transportation Research Part B*, 45(1), 267-277.
- Dai, Z., Ma, X., & Chen, X. (2019). Bus travel time modelling using GPS probe and smart card data: A probabilistic approach considering link travel time and station dwell time. *Journal of Intelligent Transportation Systems*, 23(2), 175-190.
- Degeler, V., Heydenrijk-Ottens, L., Luo, D., Oort, N., & Lint, H. (2018, July). *Unsupervised approach to public transport bunching swings formations phenomenon analysis*. Paper presented at the 14th International Conference on Advanced Systems in Public Transport (CASPT), Brisbane, Australia.
- Delgado, F., Munoz, J. C., Giesen, R., & Cipriano, A. (2009). Real-time control of buses in a transit corridor based on vehicle holding and boarding limits. *Transportation Research Record*, 2090(1), 59-67.
- Delgado, F., Munoz, J. C., & Giesen, R. (2012). How much can holding and/or limiting boarding improve transit performance? *Transportation Research Part B: Methodological*, 46(9), 1202–1217.
- Dessouky, M., Hall, R., Nowroozi, A., & Mourikas, K. (1999). Bus dispatching at timed transfer transit stations using bus tracking technology. *Transportation Research Part C: Emerging Technologies*, 7(4), 187-208.
- Dessouky, M., Hall, R., Zhang, L., & Singh, A. (2003). Real-time control of buses for schedule coordination at a terminal. *Transportation Research Part A: Policy and Practice*, 37(2), 145-164.
- Eberlein, X.J., Wilson, M.H.M., & Bernstein, D. (2001). The holding problem with real-time information available. *Transportation Science*, 35(1), 1-18.
- Estrada, M., Mensión, J., Aymamí, J. M., & Torres, L. (2016). Bus control strategies in corridors with signalized intersections. *Transportation Research Part C: Emerging Technologies*, 71, 500–520.
- Furth, P. G. (2000). Data analysis for bus planning and monitoring (No. 34). *Transportation Research*

Board.

- Furth, P. G., Hemily, B., Muller, T. and Strathman, J. G. (2003). Uses of archived AVL-APC data to improve transit performance and management: Review and potential. TCRP Web Document, 23.
- Gordon, J. B., Koutsopoulos, H. N., Wilson, N. H., & Attanucci, J. P. (2013). Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. *Transportation Research Record*, 2343, 17-24.
- Gordon, J. B., Koutsopoulos, H. N., & Wilson, N. H. (2018). Estimation of population origin–interchange–destination flows on multimodal transit networks. *Transportation Research Part C: Emerging Technologies*, 90, 350-365.
- Green, P. L., & Worden, K. (2015). Bayesian and Markov chain Monte Carlo methods for identifying nonlinear systems in the presence of uncertainty. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2051), 20140405.
- Gu, W., Li, Y., Cassidy, M. J., & Griswold, J. B. (2011). On the capacity of isolated, curbside bus stops. *Transportation Research Part B: Methodological*, 45(4), 714-723.
- Guihaire, V., & Hao, J. K. (2008). Transit network design and scheduling: A global review. *Transportation Research Part A: Policy and Practice*, 42(10), 1251-1273.
- Hadas, Y., & Ceder, A. A. (2010). Optimal coordination of public-transit vehicles using operational tactics examined by simulation. *Transportation Research Part C: Emerging Technologies*, 18(6), 879-895.
- Hans, E., Chiabaut, N., Leclercq, L., & Bertini, R. L. (2015). Real-time bus route state forecasting using partial filter and mesoscopic modeling. *Transportation Research Part C: Emerging Technologies*, 61, 121-140.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.
- Hazelton, M. L. (2001). Inference for origin–destination matrices: estimation, prediction and reconstruction. *Transportation Research Part B: Methodological*, 35(7), 667-676.
- Hazelton, M. L. (2008). Statistical inference for time varying origin–destination matrices. *Transportation Research Part B: Methodological*, 42(6), 542-552.
- Hazelton, M. L. (2010). Statistical inference for transit system origin-destination matrices. *Technometrics*, 52(2), 221-230.
- He, H., Guler, S. I., & Menendez, M. (2016). Adaptive control algorithm to provide bus priority with a pre - signal. *Transportation Research Part C: Emerging Technologies*, 64, 28–44.

- Hickman, M. (2017) Transit origin-destination estimation. In *Public transport planning with Smart card data*. Boca Raton, FL, United States: CRC Press, 15-35.
- Hounsell, N. B., & Shrestha, B. P. (2005). AVL based bus priority at traffic signals: A review of architectures and case study. *European Journal of Transport and Infrastructure Research*, 5(ARTICLE), 13-29.
- Imbens, G. (1992). An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling. *Econometrica*, 60(5), 1187-1214.
- Janos, M., & Furth, P. (2002). Bus priority with highly interruptible traffic signal control: Simulation of San Juan's Avenida Ponce de Leon. *Transportation Research Record: Journal of the Transportation Research Board*, 1811, 157–165.
- Jenelius, E. (2019). Data-Driven Metro Train Crowding Prediction Based on Real-Time Load Data. *IEEE Transactions on Intelligent Transportation Systems*.
- Jeong, R., & Rilett, R. (2004, October). Bus arrival time prediction using artificial neural network model. In *Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems* (IEEE Cat. No. 04TH8749) (pp. 988-993). IEEE.
- Ji, Y., Mishalani, R. G., & McCord, M. R. (2015). Transit passenger origin–destination flow estimation: Efficiently combining onboard survey and large automatic passenger count datasets. *Transportation Research Part C: Emerging Technologies*, 58, 178-192.
- King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9(2), 137-163.
- Koehler, L. A., & Kraus, W., Jr. (2010). Simultaneous control of traffic lights and bus departure for priority operation. *Transportation Research Part C: Emerging Technologies*, 18(3), 288–298.
- Koca, D. (2020). OD matrix estimation by deep learning using map. Bachelor thesis. Kyoto University.
- Kumar, B. A., Vanajakshi, L., & Subramanian, S. C. (2018). A hybrid model based method for bus travel time estimation. *Journal of Intelligent Transportation Systems*, 22(5), 390-406.
- Li, B. (2009). Markov models for Bayesian analysis about transit route origin–destination matrices. *Transportation Research Part B: Methodological*, 43(3), 301-310.
- Li, Y., & Cassidy, M. J. (2007). A generalized and efficient algorithm for estimating transit route ODs from passenger counts. *Transportation Research Part B: Methodological*, 41(1), 114-125.
- Li, Z., & Hensher, D. A. (2011). Crowding and public transport: A review of willingness to pay evidence and its relevance in project appraisal. *Transport Policy*, 18(6), 880-887.
- Lin, J., Wang, P., & Barnum, D. T. (2008). A quality control framework for bus schedule reliability.

- Transportation Research Part E: Logistics and Transportation Review*, 44(6), 1086-1098.
- Lin, T. M., & Wilson, N. H. (1992). Dwell time relationships for light rail systems. *Transportation Research Record*, 1361, 287-295.
- Lin, Y., Yang, X., Zou, N., & Jia, L. (2013). Real-time bus arrival time prediction: Case study for Jinan, China. *Journal of Transportation Engineering*, 139(11), 1133-1140.
- Liu, T., Ma, J., Guan, W., Song, Y., & Niu, H. (2012, June). Bus arrival time prediction based on the k-nearest neighbor method. In *2012 Fifth International Joint Conference on Computational Sciences and Optimization* (pp. 480-483). IEEE.
- Liu, Z., Yan, Y., Qu, X., & Zhang, Y. (2013). Bus stop-skipping scheme with random travel time. *Transportation Research Part C: Emerging Technologies*, 35, 46-56.
- Maalouf, M., & Trafalis, T.B. (2011). Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis*, 55(1), 168-183.
- Maher, M. J. (1983). Inferences on trip matrices from observations on link volumes: a Bayesian statistical approach. *Transportation Research Part B: Methodological*, 17(6), 435-447.
- McCord, M. R., Mishalani, R. G., Goel, P., & Strohl, B. (2010). Iterative proportional fitting procedure to determine bus route passenger origin-destination flows. *Transportation research record*, 2145(1), 59-65.
- Meng, Q., & Qu, X. (2013). Bus dwell time estimation at bus bays: A probabilistic approach. *Transportation Research Part C: Emerging Technologies*, 36, 61-71.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6), 1087-1092.
- Milkovits, M. N. (2008). Modeling the factors affecting bus stop dwell time: use of automatic passenger counting, automatic fare counting, and automatic vehicle location data. *Transportation Research Record*, 2072(1), 125-130.
- Mishalani, R. G., Ji, Y., & McCord, M. R. (2011). Empirical evaluation of the effect of onboard survey sample size on transit bus route passenger OD flow matrix estimation using APC data. *Transportation Research Record*, 2246, 64-73.
- Morales, D., Muñoz, J. C. and Gazmuri, P. (2019). A stochastic model for bus injection in an unscheduled public transport service. *Transportation Research Part C: Emerging Technologies*.
- Moreira-Matias, L., Cats, O., Gama, J., Mendes-Moreira, J., & De Sousa, J.F. (2016). An online learning approach to eliminate Bus Bunching in real-time. *Applied Soft Computing*, 47, 460-482.
- Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport

- Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, 24, 9-18.
- Muñoz, J. C., Soza-Parra, J., Didier, A., & Silva, C. (2018). Alleviating a subway bottleneck through a platform gate. *Transportation Research Part A: Policy and Practice*, 116, 446-455.
- Nesheli, M. M., & Ceder, A. A. (2014). Optimal combinations of selected tactics for public-transport transfer synchronization. *Transportation Research Part C: Emerging Technologies*, 48, 491-504.
- Newell, G.F., & Potts, R.B. (1964). Maintaining a bus schedule. *Proceedings of 2nd Australian Road Research Board*, 2, 388-393.
- Newell, G.F. (1974). Control of pairing of vehicles on a public transportation route, two vehicles, one control point. *Transportation Science*, 8(3), 248-264.
- Osuna, E.E., & Newell, G.F. (1972). Control strategies for an idealized bus system. *Transportation Science*, 6(1), 52-71.
- Patnaik, J., Chien, S., & Bladikas, A. (2004). Estimation of bus arrival times using APC data. *Journal of public transportation*, 7(1), 1.
- Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4), 557-568.
- Petit, A., Ouyang, Y., & Lei, C. (2018). Dynamic bus substitution strategy for bunching intervention. *Transportation Research Part B: Methodological*, 115, 1-16.
- Sánchez-Martínez, G. E., Koutsopoulos, H. N., & Wilson, N. H. (2016). Real-time holding control for high-frequency transit with dynamics. *Transportation Research Part B: Methodological*, 83, 1-19.
- Senevirante, P. N. (1990). Analysis of on-time performance of bus services using simulation. *Journal of Transportation Engineering*, 116(4), 517-531.
- Schmöcker, J. D., Sun, W., Fonzone, A., & Liu, R. (2016). Bus bunching along a corridor served by two lines. *Transportation Research Part B*, 93, 300-317.
- Shalaby, A., & Farhan, A. (2004). Prediction model of bus arrival and departure times using AVL and APC data. *Journal of Public Transportation*, 7(1), 3.
- Sinn, M., Yoon, J. W., Calabrese, F., & Bouillet, E. (2012, September). Predicting arrival times of buses using real-time GPS measurements. In *2012 15th International IEEE Conference on Intelligent Transportation Systems* (pp. 1227-1232). IEEE.
- Spiess, H. (1987). A maximum likelihood model for estimating origin-destination matrices. *Transportation Research Part B: Methodological*, 21(5), 395-412.

- Strathman, J. G., Kimpel, T. J., & Callas, S. (2003). Headway deviation effects on bus passenger loads: Analysis of Tri-Met's archived AVL-APC data.
- Sun, A., & Hickman, M. (2005). The real-time stop-skipping problem. *Journal of Intelligent Transportation Systems*, 9(2), 91-109.
- Sun, L., Tirachini, A., Axhausen, K. W., Erath, A., & Lee, D. H. (2014). Models of bus boarding and alighting dynamics. *Transportation Research Part A: Policy and Practice*, 69, 447-460.
- Sun, L., Lu, Y., Jin, J. G., Lee, D. H., & Axhausen, K. W. (2015). An integrated Bayesian approach for passenger flow assignment in metro networks. *Transportation Research Part C: Emerging Technologies*, 52, 116-131.
- Sun, W., & Schmöcker, J. D. (2018). Considering passenger choices and overtaking in the bus bunching problem. *Transportmetrica B*, 6(2), 151-168.
- TCRP Report 19, 1996. Guidelines for the location and design of bus stops. https://nacto.org/docs/usdg/tcrp_report_19.pdf
- Tétreault, P. R., & El-Geneidy, A. M. (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.
- Tirachini, A., Cortés, C. E., & Jara-Díaz, S. R. (2011). Optimal design and benefits of a short turning strategy for a bus corridor. *Transportation*, 38(1), 169–189.
- Tirachini, A. (2013). Bus dwell time: the effect of different fare collection systems, bus floor level and age of passengers. *Transportmetrica A: Transport Science*, 9(1), 28-49.
- Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1), 1-14.
- Trépanier, M. & Yamamoto, T. (2015). Workshop synthesis: System based passive data streams systems; smart cards, phone data, GPS. *Transportation Research Procedia*, 11, 340-349.
- Van Zuylen, H. J., & Willumsen, L. G. (1980). The most likely trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological*, 14(3), 281-293.
- Wu, W., Liu, R., & Jin, W. (2017). Modelling bus bunching and holding control with vehicle overtaking and distributed passenger boarding behaviour. *Transportation Research Part B*, 104, 175-197.
- Xuan, Y., Argote, J., & Daganzo, C.F. (2011). Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis. *Transportation Research Part B*, 45(10), 1831-

1845.

- Yu, B., Yang, Z., & Yao, B. (2006). Bus arrival time prediction using support vector machines. *Journal of Intelligent Transportation Systems*, 10(4), 151-158.
- Yu, B., Lam, W. H., & Tam, M. L. (2011). Bus arrival time prediction at bus stop with multiple routes. *Transportation Research Part C*, 19(6), 1157-1170.
- Yu, H., Chen, D., Wu, Z., Ma, X., & Wang, Y. (2016). Headway-based bus bunching prediction using transit smart card data. *Transportation Research Part C*, 72, 45-59.
- Yu, H., Wu, Z., Chen, D., & Ma, X. (2017). Probabilistic prediction of bus headway using relevance vector machine regression. *IEEE Transactions on Intelligent Transportation Systems*, 18(7), 1772-1781.
- Zhao, J., Rahbee, A., & Wilson, N.H.M. (2007). Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, 22(5), 376-387.
- Zhang, J., Shen, D., Tu, L., Zhang, F., Xu, C., Wang, Y., ... & Li, Z. (2017). A real-time passenger flow estimation and prediction method for urban bus transit systems. *IEEE Transactions on Intelligent Transportation Systems*, 18(11), 3168-3178.
- Zhang, Q., Han, B., & Li, D. (2008). Modeling and simulation of passenger alighting and boarding movement in Beijing metro stations. *Transportation Research Part C: Emerging Technologies*, 16(5), 635-649.
- Zhang, S., & Lo, H. (2018). Two-way-looking self-equalizing headway control for bus operations. *Transportation Research Part B*, 110, 280-301.
- Zhu, Y., Koutsopoulos, H. N., & Wilson, N. H. (2017). A probabilistic Passenger-to-Train Assignment Model based on automated data. *Transportation Research Part B: Methodological*, 104, 522-542.