MODELING ROUTE CHOICE BEHAVIOUR IN PUBLIC TRANSPORT NETWORK

TAN RUI

(B.Eng. (Hons.), NUS)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSIPHY

DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE

2016

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Tan Rui

05 Aug 2016

Acknowledgements

Many people have generously offered me help directly or indirectly throughout my PhD study. First of all, I would like to express my deepest gratitude to my supervisor Professor Lee Der-Horng for all the advices he has given me over the course of my PhD. He has always been a constant source of inspiration and guidance. His charming personalities with wide experience, extensive knowledge and sense of humor have not only enlightened me in my research, but also have great influence over my life.

During my research I have had the opportunity to work in the Future (Urban) Mobility IRG of Singapore-MIT Alliance for Research and Technology (SMART) under the supervision of Professor Moshe Ben-Akiva. I am deeply grateful to him for guiding me throughout the SimMobility project and my research, and giving me the opportunity to work with other researchers in SMART as well as in his ITS-lab at MIT.

I would like to express my appreciation to the members of my doctoral committee: Professor Diao Mi and Professor Meng Qiang, for their encouragement and advices me in my research. Especially, I have attended several modules lectured by Professor Meng Qiang during the PhD study. His eternal enthusiasm on teaching and research has always impressed me.

I would also like to thank all the colleaguges in Future Mobility IRG: Kakali Basak, Harish Loganathan, Milan Lovric, Joel Teo Sze Ern, Carlos Carlos, Miguel Lima Azevedo, Francisco C Pereira, Zhang Huaipeng, Wang Dong, Li Yitong, Long Wei Ling Janet, Andrew Tong Kwok Cheong, Cecille Maquito, Zuo Bingran, etc. I own special thanks to the following modelers that I have

worked with together on modeling passenger route choice behaviour: Steve Robinson, Sebastian Raveau Feliu, Adnan Muhammad, Lu Yang, and Carlos Carrion.

Thanks to all the colleagues and friends in NUS for their support and accompany: Li Siyu, Sun Lijun, Zhao Kangjia, Katarzyna Anna Marczuk, Hao Siyu, Jin Jiangang, Wu Xian, Lu Zhaoyang, He Nanxi, Wang Yadong, Xu Min, Hua Wen, Fu Rao, Qin Han, Foo Chee Kiong, Charulatha Vengadiswaran, etc.

Finally, I would especially like to thank my parents, my parents-in-law and my husband Li Xiao for always giving me their unconditional love, support and understanding. Without their constant support and encouragement, I could not finish this thesis.

Table of Content

List of Tables

List of Tables

List of Figures

Summary

In an urban city like Singapore, the majority of travels are conducted by public transport services especially with the increasing travel demand and limited land resource. Public transport ridership has been continuously increasing since past decades. Passengers' travel behaviour is highly correlated with the performance of public transport systems. It is therefore essential to identify relevant factors that affect passengers' travel behaviour in a realistic manner. This thesis focuses on the pre-trip route choice behaviour of passengers in the multimodal public transport network of Singapore. For the analysis of the problem, discrete choice models and disaggregate revealed preference data are adopted.

Empirically, this thesis identifies relevant factors that affects passenger route choice behaviour in public transport network using revealed preference data from smart card and surveys. Methodologically, it formulates a new path-size definition to address the correlation caused by path-overlapping with special consideration to the unique characteristics of public transport network in dense urban cities, and it proposes a latent class route choice model framework to address the availability issues of different access/egress modes in the multimodal public transport network. Practically, the application on the rapid transit network in Singapore has translational impact on the prevailing travel data as it complements the missing transfer station in smart card data. It demonstrates how to apply route choice model and get realistic estimation of passenger flows on rapid transit stations and lines.

Chapter 1 - Introduction

1.1 Research Motivation

Public transport is by far the most efficient mode of transport, in terms of both land usage and energy consumption. It is effective in reducing congestion and environmental pollution. In large urban cities with high population density like Singapore, majority of the travels are conducted by public transport due to the increasing travel demand and limited land resource. Singapore has devoted lots of efforts to build its world-class transport system, and is continuously taking initiatives to improve/enhance the system by further integrating land-use and town design. Due to the extreme land scarcity and high population density in Singapore, public transport is and will always be the major transport mode.

Singapore's public transport covers a comprehensive range of mobility services, with mass rapid transit (MRT) serving heavy traffic corridors, Light Rail Transit (LRT) complementing MRT as the feeder services, and buses widely covering entire residential areas. According to the latest Household Interview Travel Survey (HITS) conducted by Singapore Land Transport Authority (LTA) in 2012, the public transport mode share in Singapore has risen up from 59% in 2008 to 63% in 2012 (Singapore Land Transport Authority, 2013). Public transport ridership has been grown by 3.7% to an average 6.36 million trips per day in 2013, being the ninth consecutive rise since 2005 (The Straits Times, 2014). Despite of the high efficiency of the public transport network, challenges remains to further improve the public transport ridership and reduce the car population by providing more attractive multimodal public transport services.

Unlike car trips on road network, public transport trip could involve not only transfers among different travel modes in public transport systems, but also access from origin to public transport services and egress from public transport services to destination (Ortúzar and Willumsen, 2001). Passengers' travel behaviour is highly correlated with the performance of public transport systems. By understanding how passenger select their routes, route choice models are extensively used to evaluate transportation network performance, assess policies' effectiveness, appraise various intelligent transportation systems, and predict travel activities in future scenarios (Prato, 2009). It is therefore essential to understand passengers' route choice behaviour in public transport networks in a realistic manner.

However, existing literature on route choice modelling has been primarily concentrated on the behaviour of car drivers in road networks rather than passenger behaviour in public transport networks (Anderson et al., 2014). To fill up this gap, this thesis is dedicated to the modeling of passengers' route choice behaviour in public transport network, with application to Singapore.

1.2 Research Objectives and Scopes

The focus of the thesis is on the pre-trip route choice behaviour of passengers in public transport networks. For the analysis of the problem, discrete choice models and disaggregate revealed preference data are adopted based on Singapore's real public transport network. Efforts are made to get more accurate representation of route choice behaviour, proper treatment to address correlation among alternatives and estimation of behaviour parameters against real data.

The core objectives of this thesis are served through Chapter 3 to Chapter 6 with several specific issues to address.

The first part of the thesis identifies and quantifies different aspects of travelling that affect passengers' route choice decisions from initial boarding stop to final alighting stop using smart card data in Chapter 3. It empirically evaluates six choice set generation approaches to the multimodal public transport network in Singapore under a comprehensive evaluation framework. A stop-to-stop route choice model is estimated based on final choice set with 100% coverage against one data set and the model is validated by examining the prediction performance against another two datasets.

Chapter 4 is devoted to proper treatment to address correlation among alternatives due to path overlapping. A new path-size formulation is proposed is presented to address the path-overlapping issue with special consideration of the unique characteristics of dense public transport network in urban cities. This formulation is comparatively analyzed and validated against different route choice models.

The third part of the thesis aims at modelling a complete public transport route choice from origin building to destination building with multimodal access and egress using survey data. This study expands existing literature by considering the multimodality of public transport trips, not only among different public transport services, but also among different access/egress modes. Six possible access/egress modes are considered including walk, bicycle, taxi, car as driver, car as passenger and motorcycle, while all public transport services in Singapore are included. While walking and taxi can be assumed to be available to all

passengers, this might not be the case with other access/egress modes, particularly for access/egress in car as passenger. A latent class route choice modelling framework is proposed to address this availability issue in modelling passenger route choice decisions. The latent class model utilize a Logit model structure to examine the availability of access/egress modes depending on the passenger social-economic characteristics and trip characteristics. Conditionally on the choice set with specified modal availability in the latent class, the route choice decision is modelled considering both path attributes and passengers socio-economic characteristics. Data from household travel surveys is used to unveil passengers' route choice preferences in the public transport network.

The last part of the thesis presents an application of route choice modeling to assess the performance of the rapid transit network in Singapore by combing both smart card and survey data. Smart card data in Singapore is comprehensive in terms of passenger demand coverage, but it is incomplete as there is no record on transfer stations. The household travel survey data in Singapore stores complete information of trips but it is only collected on a small fraction of passengers. This work first models passengers' route choice decisions on the rapid transit network using travel survey data and then estimates the passenger boarding demand, alighting demand and transfer demand given the complete travel demand recorded in smart card data. It assess the performance of rapid transit network in Singapore in terms of passenger flows on transit service lines, transfer demand at stations, and probability of fail-to-board and fail-to-seat at important transfer station.

1.3 Thesis Organization

This thesis is organized as follows:

- Chapter 1 introduce the concept of route choice modeling, research objectives and scopes.
- Chapter 2 reviews the related work on route choice modeling in literature, focusing especially the models and applications on public transport networks.
- Chapter 3 identifies and quantifies different aspects of travelling that affect passengers' route choice decisions from initial boarding stop to final alighting stop using smart card data
- Chapter 3 proposes a dedicated path-size formulation to address path overlapping in public transport network.
- Chapter 5 models passengers' route choice decisions from origin building to destination building with multimodal access/egress.
- Chapter 6 applies a dedicated route choice model to the rapid transit network in Singapore and evaluate the network performance by combining smart card data and survey data.
- Chapter 7 concludes the completed work and presents directions for future research.

Chapter 2 - Literature Review

In this chapter, the state of the art route choice models and applications in public transport network is presented. Firstly, section 2.1 presents the overview of route choice modeling, while Section 2.2 reviews data used for route choice models on public transport network. In Section 2.3, literature review on network representation for public transport network is presented, followed by review on choice set generation approaches and evaluation in section 2.4. Last, in section 2.5, route choice models and their applications in public transport networks are presented.

2.1 Route Choice Modeling Overview

Given an origin-destination pair in a transport network, route choice models assess passengers' perception on various path attributes such as in-vehicle travel time, waiting time, walking time, number of transfers, and etc. Route choice decisions are also affected by passengers' socio-economic characteristics such as gender, age, income and trip purpose. Modeling route choice decisions in dense urban public transport network is particularly a challenging problem. On one hand, the high density of urban public transport network imposes operational constraints on route choice models for large-scale network. On the other hand, the number of possible paths increases in a combinatorial dimension.

The operational constraints of large-scale network lead to a simple deterministic route choice model – all or nothing approach. The key assumption in this approach is that travelers only choose the least cost paths from origin to destination. Here the cost of a path is generally defined as an additive cost of

each link along the path. It has been widely applied in most of the traffic assignment models in which estimating realistic travel behaviour from real data is not of key interest. Despite the cost function could be defined differently, this approach is not able to capture the unobserved or unknown uncertainty that influences travelers' decisions.

The widely used discrete choice models, belonging to the family of random utility model framework, are particularly suitable to model choice behaviour with uncertainty including, but not limited to, household activity choice, destination choice, mode choice and route choice (McFadden, 2000). Taking route choice as an example, under this framework, travelers are assumed to evaluate each path alternative from origin to destination and choose the one that maximizes his/her utility out of a set of discrete path alternatives. The path utility depends on three important aspects. The first aspect represents influential factors that affect how passengers value a path. Influential factors not only include path attributes but also traveler's socio-economic characteristics, as well as the choice situation. The second aspect refers to preference parameters that capture passengers' preference on different influential factors. These unknown preference parameters can be identified through optimization techniques against real route choice data. The third aspect is the random term that is assumed to associate with each path utility. The inclusion of this random term has brought uncertainty into choice-making process and it influences the probability of an alternative being chosen from a set of alternatives.

To achieve more realistic travel behaviour models, researchers recently focused on three major issues: accurate representation of actual behaviour, proper

treatment to address correlation between alternatives and estimation of behaviour parameters against real data. Frejinger (2008) presents a schematic overview of modeling route choices in this approach, as depicted in *Figure 2-1*.

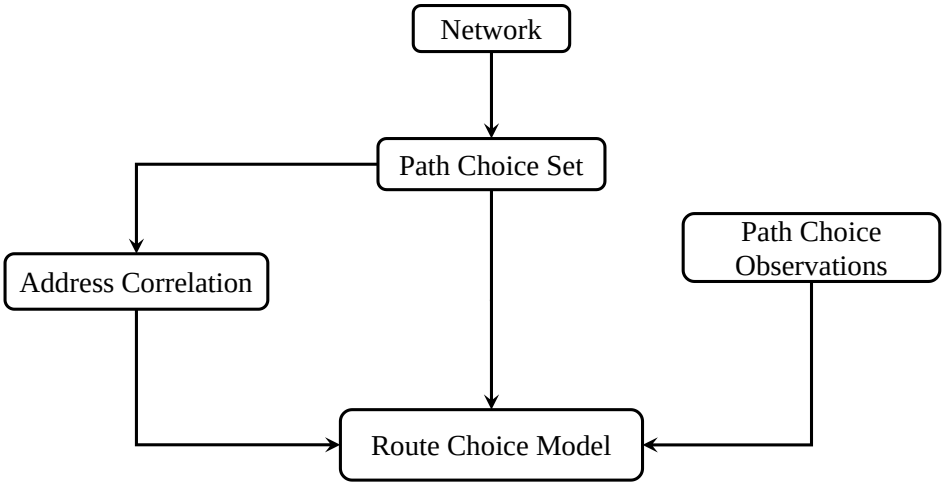


Figure 2-1: Overview of route choice modeling

Path choice behaviour observations can be obtained by conducting traditional travel surveys or through new ITS technologies such as Global Positioning System (GPS) for vehicles and smart card for passengers. As discrete choice model deals with disaggregated choice of individual, it is necessary to generate a path choice set for each path observation. Therefore, before estimating the parameters involved in a route choice model, the set of possible alternative paths are firstly generated explicitly for the examined network. This process is called choice set generation. In a dense urban public transport network, it is far from a trivial task as the number of feasible path alternatives for a given origin-destination pair is huge. For a dense urban public transport network, path alternatives are highly correlated due to high degree of overlapping among paths. Proper handling of the correlation among alternatives is therefore essential.

Finally, in the route choice model, the utility function of path alternatives has to be properly specified and behaviour parameters will be estimated against real route choice observations.

2.2 Route Choice Data

Actual route choice data for public transport passengers is of great importance when assessing generated choice sets and estimating route choice models. However, rather few studies of passengers' route choice behaviour on public transport networks using revealed preference data are available in the literature. One major reason is that such data is difficult to collect. It needs a lot of information to describe passenger selected path in public transport network, as the trip on public transport network is always multimodal with transfers between different services. Route choice data on public transport network can be collected by either asking passengers to describe their selected path through surveys, or by passive monitoring them using smart card system.

Most of the available literature on modeling passengers' route choice behaviour utilize travel survey data. Few studies on rail passenger route choice behaviour use route choice data collected for train users via face-to-face interviews in a specific train corridor in the Netherlands (Hoogendoorn-Lanser and Bovy, 2007; Uges et al., 2002). Anderson et al. (2014) conducted route choice survey both online and via telephone for the multimodal network of the Greater Copenhagen. The public transport modes considered in their work include bicycle, walking, bus, train and metro. Some other studies focus on modeling route choice behaviour on metro network only based on travel survey data collected at metro stations (Guo and Wilson, 2011; Raveau et al., 2011). Clifton and Muhs (2012)

presented a recent review on various approaches to collect route choice data on multimodal transport network though travel survey.

There are very limited applications of smart card data for modeling route choice behaviour from automated fare collection system on public transport network. In cities that smart cards have to be tapped in and tapped out both at boarding of each new service line and at alighting of each service line, passengers' route selection on public transport network is directly available with comprehensive coverage. Sun and Xu (2012) focus on analysis of travel time reliability and estimation of passenger route choice behaviour on the Metro network in Beijing using smart card data. (Schmöcker et al., 2013) used the route choice data in smart card to validate the generation of choice set on public transport network. Kusakabe et al. (2010) apply smart card data to study passengers' train choices made by railway passengers in the railway network of Japan. (Jánošíková et al., 2014; Nassir et al., 2015) estimated stop-to-stop route choice models using smart card data.

Comparing to survey data, the major drawback of smart card data is that it only collects route choice data on the use of public transport services and there is no information on access from trip origin and egress to trip destination. Besides, in general, there lacks information on passenger socio-economic characteristics. But the high spatial and temporal coverage of trips at high data accuracy of smart card data still makes it attractive for passenger route choice modeling. Trepanier et al. (2009) compare household travel survey data with smart card data. They showed that the smart card data is partially consistent with the travel

survey data and it is applicable to analyze transit behaviour if the insufficient parts of the data are supplemented or negligible.

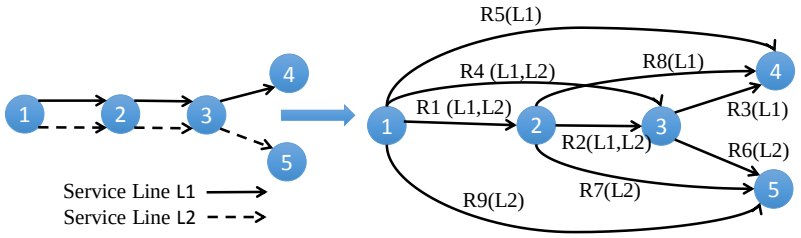
2.3 Representations of Public Transport Network

Public Transport Network formulation for passenger route choice models can be categorized into two groups: frequency-based network formulation and schedule-based network formulation. In frequency-based network, service lines are assumed to operate at deterministic or dynamic frequency, whereas in schedule-based network, each run of each service line is considered explicitly according to timetable or real arrival/departure time. The following two sections summarize the frequency-based and schedule-based approaches for public transport network formulation available in literature respectively.

2.3.1 Frequency-Based Approach

Based on a simplified test network with few service lines serving several stations, Dial (1967) proposes one of the first frequency-based transit assignment models. The model is based on a normal network with travel time as link attributes, and waiting time as node attributes. Travel time is assumed to be the same for all links while waiting time is modeled as half the headway of all service line leaving the node. There are two major drawbacks of Dial's work. Firstly, this network does not handle situations where service lines connect the same station pairs have different travel distance or travel time. Secondly, Dial does not address common line problem that travelers face multiple service line choices from his current station to the next transfer station and passengers prefers to board the first arriving service in order to minimize total travel time (Chriqui and Robillard, 1975).

A "route section" model was proposed by De Cea and Fernández (1993) with reference to the dual network representation based on Chriqui and Robillard (1975). The idea is that each stop-to-stop pair that can be traversed without change of service line is represented as a separate link, called "route section" or "route segments", as illustrated in Figure 2-2. They assume that passenger will always board the first arriving vehicle of interest from the predetermined attractive service lines that leads to their destination. The expected travel time on each "route section" is then modeled as a function with respect to all service line frequencies in that "route section". As there is no change of service line within each "route section", waiting time can be considered more precisely at the boarding node of a "route section" only.



Rx(L*i*, L*j*) : Route Section x transformed from service line L*i*, and L*j*

Figure 2-2: Illustration of network formulation used in De Cea and Fernández (1993)

The major advantages of this network presentation are: 1) it addresses the common line issue in public transport network as all common lines serving the same stop pair are embedded in the same route sections and therefore, the effect due to common lines between stop pairs can be compensated in waiting time to embark on each route segment (De Cea and Fernández, 1993); 2) it should help to speed up the shortest path searching as route segments are created between

stop pairs as long as they can be reached via at least one service line without a transfer, regardless of their actual distance, and 3) aggregation of common service lines on route segment gives freedom to model the boarding behaviour and interactions between passengers and service vehicles in an agent-based simulation environment. The disadvantage of this approach is the number of route segments created increases quadratically with the number of stops along each service line. This requires larger memories for data storage.

Under the same behaviour assumption, Nguyen and Pallottino (1988) introduce the concept of hyper-path in network formulation. A hyper-path is defined to be a sequence of paths sets connecting origin and destination. Each path set contains a set of paths that could be taken by passengers from station to station along the hyper-path and path represents part of service line associated with hyper-path. The probability of choosing a particular path out of a path set depends on the line frequencies. Similarly, Spiess and Florian (1989) illustrate this problem by introducing a concept called strategies that passengers adopt to predetermine a set of attractive service lines that lead to their destination and assume that passenger will always board the first arrival vehicle from the attractive set. Their model presents some important limitations, the major one being that waiting time at transfer stops is not affected by transit demand, neglecting the congestion effect that rises under capacity constraints.

In recent studies, various advanced frequency-based route choice models have been developed to address the effect of capacity constraints on passengers' route choice behaviour. Lam et al. (2002) proposed a network formulation approach with elastic line frequency under capacity constraints, in which the line

frequency is modeled as a function of passenger flows and line capacity. Schmöcker et al. (2008) propose the first dynamic frequency based network formulation for public transport networks by introducing "fail-to-board" probability which represents the crowding effect that in some circumstances, passenger is not able or willing to board the first arriving service. This is a first approach to dynamic frequency-based transit assignment. The network was constructed on a hyper-path, the cost of which is subject to fail-to-board probability. The approach is illustrated in a simplified test network and then applied to London Underground Network. Schmöcker et al. (2011) further incorporate "fail-to-seat" probability into their frequency-based hyper-path formulation. A similar concept of "fail-to-board" and "fail-to-sit" has been adopted by Leurent (2008). In his model, however, new arcs are added from each possible boarding station to each possible alighting station of each service line. Standing arcs and sit arcs are also added to represent different utility associated with standing and seating. Leurent (2009) further apply this approach on Paris network and claim that seat availability has a significant impact on perceived utility and therefore affect the line loadings.

Despite the frequency-based approach with hyper-path concepts has been used by some researchers in route choice models, they are widely used in traffic assignment approach where route choice behaviour is not carefully extracted from real route choice observations. To the author's knowledge, analysis of passengers' hyper-path choice behaviour or path strategy choice is not available in the literature yet. There lacks of data recording passengers' path strategy choices, which becomes the major hinders to this approach in analyzing choice behaviour. Besides, how to derive proper transport economics factors such as

value of time, rate of substitution from hyper-path choice behaviour still remains a challenging question.

2.3.2 Schedule-Based Approach

Schedule-based approach based on time-expanded network representation have been proposed for public transport systems in order to consider more coherent user behavioural hypotheses in relation to service characteristics since 1990s, especially for sub-urban networks or inter-city rail networks. The main advantages of time-expanded network representation are firstly, standard shortest path algorithm can be applied directly on the time-expanded graph, and secondly, congestion can be considered as increase in link impedance directly. Nuzzolo and Russo (1996) represent one of the earliest time-expanded schedule-based networks to handle low frequency services. They built a time-expanded network model, also called diachronic network, to represent each run of each service line in both time and space dimension. The diachronic graph consists of two sub-graphs: (1) a service sub-graph representing run-based transit service lines according to its scheduled arrival time at each stop; and (2) a demand sub-graph discretizing travel demand onto time segmented nodes. On the service sub-graph, each station node is expanded by adding associated arrival node and departure node. Alighting link is added between each arrival nodes its associated stop node, and stop node is connected to its departure node via departure link. Nuzzolo et al. (2001) further modified this approach by formulating a diachronic graph consisting three sub-graphs: (1) a service sub-graph (2) a demand sub-graph and (3) an access/egress sub-graph linking up demand sub-graph with service sub-graph, as well as creating possible transfers

between stops within service sub-graph. The service sub-graph is illustrated in Figure 2-3. They also applied the same network formulation approach to consider capacity constraints explicitly (Nuzzolo et al., 2012). Sumalee et al. (2009) represents a modified time-expanded transit service network based on the diachronic graph approach proposed by Nuzzolo et al. (2001) to incorporate dwelling time. Dwelling links representing the dwelling time of a vehicle at a stop are newly added, creating temporal linkage for each stop

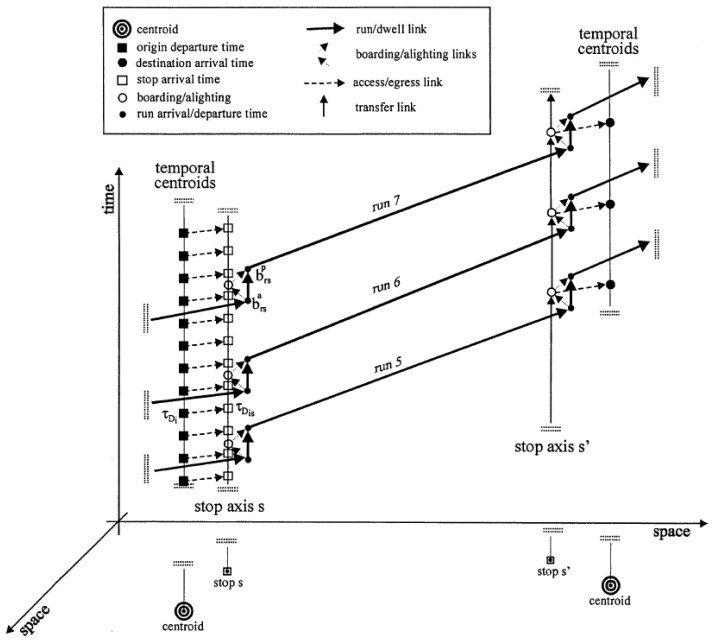


Figure 2-3: A diachronic graph representation for schedule-based network (Source: (Nuzzolo et al., 2001))

Hamdouch and Lawphongpanich (2008) present some extensions to complete the conventional time-expanded schedule-based network by: (1) adding walking links to allow passenger transferring between two nearby stations, (2) creating bicycling link and driving link in parallel to walking link for access, (3) generating waiting links to represent waiting at the station for arrival vehicle. Each link is associated with a capacity constraint where walking link have an

infinite capacity, and travel link has capacity in line with its service line or vehicle. Each node is also assigned a capacity cap according to the size of platform. First come first serve strategy is adopted during traffic simulation. Most recent works such as (Hamdouch et al., 2011; Zhang et al., 2010) also adopts time-expanded schedule-based network for public transport route choice modeling.

Comparing to schedule-based approach, frequency-based approach is more appropriate to dense urban public transport network with high frequency services. It requires less detailed input data, and therefore reduces computational complexity. The common lines problem can be handled easily by frequency aggregation in frequency-based network. However, path utility has to be updated dynamically in order to reflect the network conditions in real time. While for schedule-based approach, it is more suitable for low-frequent services. As stated by Nguyen et al. (2001), the inherent advantage of schedule-based networks is that they allow dynamic network description, time-expanded minimum path searching, and dynamic network loading procedure.

2.4 Choice Set Generation Methods

Only a limited number of papers describe the use of choice set generation approaches in public transport networks. Compared to road networks, choice set generation for multimodal public transport network is more complicated and challenging because public services only operate at fixed stops along fixed routes with scheduled timetables/frequencies. It therefore generates the need for waiting and transfers. Applications of choice set generation approaches in public transport networks are mainly based on choice set generation approaches

used in road networks with implicit or explicit modifications to account for the discontinuous serviceability and connectivity. Therefore, this section reviews the choice set generation approaches on road networks first, followed by choice set generation approaches in public transport networks. Literature evaluating choice set generation approaches are also presented.

Note that there are also various studies using sampling of alternatives to get a choice set with consideration to a full choice set in discrete choice models (Flötteröd and Bierlaire, 2013; Frejinger et al., 2009; Lai and Bierlaire, 2015; Li et al., 2016) or recurving generating subsequent link choice to avoid generating choice set with path alternatives (Fosgerau et al., 2013; Mai, 2016; Mai et al., 2015). However, the former approach depends on an available full choice set which covers all possible alternatives while this full choice set is not directly available and extremely hard to enumerate in the route choice context, especially for large transport network, while the latter approach is still undergoing a method development stage and is not necessarily mature enough for a reliable usage in this context for behavioral analysis.

2.4.1 Choice Set Generation for Road Network

In general, current choice set generation approaches for explicitly generating path choice sets can be categorized into two groups: deterministic choice set generation approaches and stochastic choice set generation approaches.

In deterministic choice set generation approaches, network attributes and other parameters remain deterministic. Choice sets are generated by iteratively searching for the least cost path. Several approaches are presented in the

literature to generate choice set deterministically. Link penalty approach and link elimination approaches both iteratively search for a new shortest path after updating the link cost on certain links. The link penalty approach increases all link costs along the current shortest path gradually (de la Barra et al., 1993). In the link elimination approach, one or more links on the shortest path/paths are removed, or the cost set to infinite, before searching for a new shortest path (Azevedo et al., 1993; Rieser-Schüssler et al., 2012). The k-shortest path algorithm generates the first k shortest loop-less paths from an origin to a destination in the network using various well-structured link elimination heuristics (Van der Zijpp and Fiorenzo Catalano, 2005; Yen, 1971).

The aforementioned three approaches are all based on a fixed cost function during each path search. In order to address travelers' heterogeneity on path cost definition, Ben-Akiva et al. (1984) proposed a labeling approach by using different cost functions, called labels, to identify different "shortest" paths. The branch & bound technique is applied to path generation as selective enumeration with specified constraints as a search bound. Various constraints on path feasibility, path logicality, path compositions, and travelers' behaviour can be applied. All feasible path alternatives satisfying the constraints are generated between origin and destination (Prato and Bekhor, 2006).

Stochastic choice set generation employs stochasticity in path searching from two approaches. One common approach, called the simulation approach, is to randomize network attributes from a probability distribution. During each round of least cost path search, the link cost in the network is drawn from a pre-specified distribution with respect to its link cost, such as normal distribution

(Bekhor et al., 2006; Ramming, 2001), and truncated normal distribution (Carlier et al., 2003; Nielsen, 2000; Prato and Bekhor, 2007). Another stochastic approach is called the "Random Walk" approach. It introduces stochasticity into path-establishing procedure by successively selecting the next link/node based on a certain distribution. Starting from the origin node, the path is constructed by successively selecting the next link based on a Kumaraswamy distribution, depending on the link distance and subsequent path distance from the link to the destination (Frejinger et al., 2009).

2.4.2 Choice Set Generation for Multimodal Network

In multimodal networks, separate single mode networks are usually merged into a single integrated network by interconnecting different service modes via feasible walking links, and representing temporally discontinuous services as waiting links or waiting attributes on transit links. Conventional choice set generation approaches can then be directly applied on the integrated multi-modal network to find path alternatives between any origin and destination. The k-shortest path algorithm and multi-objective shortest paths approach were applied to a multimodal network by Abdelghany and Mahmassani (1999). Florian (2004) also adopts a multi-objective shortest paths algorithm considering the dominancy of both time and cost based on Dijkstra's label-setting algorithm. Although Abdelghany and Mahmassani (1999) claimed that multi-objective shortest paths approach outperforms k-shortest paths approach in terms of coverage, which is defined as the percentage of recorded alternatives being available in the generated choice set, multi-objective shortest paths

approach is not applicable to large-scale networks as its computational complexity is well-known to be NP-hard with respect to network size.

Benjamins et al. (2001) applied simulation approach to generate choice sets for different traveler groups in a multimodal network, with both private and public networks, using a normal distribution. Fiorenzo-Catalano et al. (2004) extends this approach on multimodal networks by introducing stochasticity into both link cost and travelers' preferences. Travelers' behavioural parameters are firstly randomized to generate different cost functions, following which, the link costs are randomized resulting in different paths being selected. Friedrich et al. (2001) applied the branch and bound algorithm for generating path choice sets on a public transport network by converting the network into a time-dependent network. Hoogendoorn-Lanser et al. (2007) extended Friedrich's approach on a multi-modal network, with both a public transport and road network.

2.4.3 Assessment of Choice Set Generation Methods

As the quality of path choice set strongly influences the subsequent route choice model estimation, it is essential to evaluate the generated choice set by different approaches. However, literature on evaluating different choice set generation algorithms is rare. Only a few papers which evaluated various choice set generation approaches on the road network were found. These are described below.

Bekhor et al. (2006) evaluated four choice set generation algorithms: labeling, link elimination, k-shortest paths and simulation on road network in Boston, U.S.. Coverage, computational time, cumulative distribution of choice set size and cumulative distribution of the number of links in the choice set were used

to evaluate the algorithms. This work was performed on a highway network in Boston, U.S., consisting of 13,000 nodes and 34,000 links, while the evaluation data was from a university transportation survey with 188 respondents in total. The coverage, which was defined as the percentage of recorded alternatives being available in the generated choice set, was reported to be 72% for the labeling approach with16 labels, 60% for link elimination approach, 57% for k-shortest paths approach with k equals to 40, 50% for simulation approach at 48 draws, and 84% for the combination of all approaches. The fastest computational time for one origin-destination (OD) pair was labeling approach at 32 seconds, followed by simulation approach and link elimination approach at a few minutes. k-shortest paths approach was reported to take more than one hour.

The recent work presented by Rieser-Schüssler et al. (2012) is probably the first available literature evaluating different choice set generation approaches for large-scale road network using car trips extracted from GPS data. They comprehensively compared link elimination approach and simulation approach, by evaluating the following: the coverage, the average path size in the choice set indicating the degrees of overlapping among alternatives, the path distance variation in the choice set, computational performance in terms of route set size, the least cost distance, the number of non-pass nodes on the least cost path, and the time abort threshold which is defined as the cut-off computational time for generating paths for each OD pair. The evaluation data set contains around 36,000 car trips with 2,434 ODs in Zurich, Switzerland using a Navteq road network with 408,636 nodes and 882,120 links. The link elimination approach achieves 73% coverage with a choice set size of 100 within 10 minutes, while

the simulation approach covers 75% of the reported trips at a computational time of over an hour.

There is currently no literature available on evaluating different choice set generation algorithms on multimodal public transport network.

2.5 Route Choice Models

In this section, an overview of route choice models in general is presented. The merits and drawbacks of each model will be discussed and the applications of these models in public transport networks will be particularly presented. Other literature reviews on route choice models are available. Ben-Akiva and Bierlaire (2003) present a brief summary of using discrete choice models for route choice modeling. Frejinger (2008) presents route choice analysis in a static road network. A recent literature review by Prato (2009) reviewed the analysis of route choice behaviour under the framework of discrete choice models comprehensively. Two useful books on discrete choice models are highly recommended (Ben-Akiva and Lerman, 1985; Train, 2009).

Discrete choice models can be classified based on different model structures. In this review, the simple Multinomial (MNL) logit model and its modification, Path-Size Logit that maintains the same logit structure by adding a correction term to the deterministic part of the utility function will be firstly introduced, following by Multimonimal Probit Model that captures correlation between alternatives in the covariance matrix of the error term in the utility function. Generalized Extreme Value (GEV) models which support multi-dimensional structure of the choice set will also be introduced, as well as Mixed logit models which capture heterogeneity.

2.5.1 Multinomial Logit

The Multinomial Logit (MNL) model is basic and the simplest logit, which has the closed form of selection probability under the assumption that the error terms are identically and independently distributed (i.i.d.). In MNL, the probability of choosing path i with path utility V_i in choice set C_n is expressed as:

$$P(i|C_n) = \frac{exp(V_i)}{\sum_{j \in C_n} exp(V_j)} \qquad \text{Eq.(2-1)}$$

MNL model is most commonly used in practice due to its simplicity. (Jánošíková et al., 2014) presented a stop-to-stop route choice model estimation on public transport network using MNL model based on smart card data. In this model, in-vehicle time, transfer walk time, number of transfers and waiting time were considered, while the choice set was directly inferred from smart card observations. (Nassir et al., 2015) analyzed the route choice behaviors from smart card data in Brisbane using binary logistic model, which is an MNL model with binary options in choice set.

However, the assumption that the error terms are i.i.d. in MNL model does not valid in the context of route choice, particularly due to paths overlapping. Efforts have therefore been made to overcome this restriction by making a deterministic correction of the utility for overlapping paths. Given the shortcomings of the MNL model, more complicated models have therefore been proposed in the literature to overcome this issue. One easy and practical way is to add a deterministic correction term to the path utility function while maintaining the simple path selection structure of MNL.

2.5.2 Path Size Logit

Cascetta et al. (1996) proposed the first deterministic correction to address the correlation between alternatives due to path overlapping. They introduced an attributed, called "Commonality Factor" (CF) with three types of different formulations to path utility function. However, the formulations of the CF attribute were provided without any theoretical proof on such formulations or which of the formulation should be used.

Later, Ben-Akiva and Ramming (1998) proposed the Path-Size Logit (PSL) model in which a deterministic correction attribute "path size"(PS) was introduced and added to the path utility $U_{in} = V_{in} + \beta_{PS} ln(PS_{in}) + \varepsilon_{in}$ where V_{in} is the deterministic path utility, ε_{in} is the error component, and the correction attribute: path-size is defined as:

$$PS_{in} = \sum_{a \in \Gamma_i} \left(\frac{l_a}{L_i}\right) \frac{1}{\sum_{j \in C_n} \delta_{aj}} \qquad \text{Eq.(2-2)}$$

Where PS_{in} is the path-size of path i in choice set C_n; $l_a$is the length of link a , Γ_i is the set of all links along path i; L_i is the total length of path i; δ_{aj} equals to 1 if link a is on path j and 0 otherwise. Although the idea of path-size is similar to the "Commonality Factor", the derivation of path-size formulation was provided based on discrete choice theory for aggregate alternatives.

Another formulation of path-size was presented by Ben-Akiva and Bierlaire (Ben-Akiva and Bierlaire, 1999) by considering the length of the shortest path in the choice set, $L^*_{C_n}$,:

$$PS_{in} = \sum_{a \in \Gamma_i} \left(\frac{l_a}{L_i}\right) \frac{1}{\sum_{j \in C_n} \delta_{aj} \frac{L^*_{C_n}}{L_j}} \qquad \text{Eq.(2-3)}$$

The major drawback of this formulation is that it will produce path-size value larger than 1 for unique path alternatives that are longer than the shortest path, which will result indifferent signs for the logarithm of the path-size in the utility function.

Late, a generalized path-size formulation by adding a non-negative scaling parameter γ to *Eq.(2-3)*(Ramming, 2001)

$$PS_{in} = \sum_{a \in \Gamma_i} \left(\frac{l_a}{L_i}\right) \frac{1}{\sum_{j \in C_n} \delta_{aj} \left(\frac{L_i}{L_j}\right)^{\gamma}} \qquad \text{Eq.(2-4)}$$

Note that when γ equals to 1, it is similar to the path-size formulation as in *Eq.(2-3)*. Numerical demonstration has shown that this formulation has a better estimation result with γ set to infinity than the model with γ equals to 1. With large γ, this formulation can help to decrease the impact of extremely long paths in the choice set. For example, when a short path shares a link with an extreme long path, the path-size for the short path will have a positive impact on the utility while the PS for the long path will have a negative impact on the utility. Estimation results of C-Logit and PSL indicated that PSL with generalized path-size formulation outperforms C-Logit in terms of model estimation (Ramming, 2001).

Frejinger and Bierlaire (Frejinger and Bierlaire, 2007) proposed a framework to combine deterministic attribute of path-size as in PSL and stochastic attribute of path overlapping in Error Component models. For the deterministic path-size,

they presented derivation of the path-size formulation as in *Eq.(2-3)*, and provided analytical and numerical evaluation on the generalized path-size formulation as in *Eq.(2-4)*. Their results concluded that the generalized path-size formulation might produce counter intuitive results and the original formulation is more appropriate.

The basic path-size formulation in *Eq.(2-3)* has been extended by Bovy et al. (2008). They provided another derivation of path-size formulation based on nested Logit model. Based on the derivation, they further proposed a path size correction factor as follows:

$$PSC_{in} = \sum_{a \in \Gamma_i} \left(\frac{l_a}{L_i}\right) ln \sum\nolimits_{j \in C_n} \delta_{aj} \qquad \text{Eq.(2-5)}$$

where PSC_{in} denotes the path-size correction factor of path $i$in choice set C_n. Note that unlike the path-size formulation in *Eq.(2-3)*, the logarithm is taken inside the path-size correction factor, and therefore the utility function is $U_{in} = V_{in} - \beta_{PS} PSC_{in} + \varepsilon_{in}$.This formation is theoretically more appealing, but both the estimation and prediction results show similar performance to the path-size formulation in *Eq.(2-3)*.

Frejinger et al. (2009) presented an extended path-size formulation based on the formulation in *Eq.(2-3)* by including an expansion factor:

$$EPS_{in} = \sum_{a \in \Gamma_i} \left(\frac{l_a}{L_i}\right) \frac{1}{\sum_{j \in C_n} \delta_{aj} \, \Phi_{jn}} \qquad \text{Eq.(2-6)}$$

Where Φ_{jn} is the expansion factor defined as:

$$\Phi_{jn} = \begin{cases} 1 \;\; if\, \delta_{aj} = 1 \; or \; q(j)R_n \geq 1 \\ \dfrac{1}{q(j)R_n} \qquad otherwise \end{cases} \qquad \text{Eq.(2-7)}$$

Where $q(j)$ is the probability of path j being selected by sampling from universal choice set and R_n is the total number of sampled path in choice set C_n. They presented numerical results based on synthetic data and concluded that this expanded path-size is remarkably better than original path-size formulation as unbiased parameter estimates can be obtained. However, the illustration is only performed using synthetic data on a small toy network. Computing the expansion factor is expected to be extremely intensive in large-scale network where the universal choice set is numerous.

Most of the public transport route choice models either neglect the path-overlapping issue, or directly adopt the path-size formulation in PSL model or C-Logit model without explicitly defining it in the public transport context (Anderson, 2013; de Grange et al., 2012; Guo, 2011). They are only two such studies exist that addressed the path-size formulation for route choice models in public transport networks, which are reviewed below.

Hoogendoorn-Lanser et al. (2005) implemented a PSL model to explicitly address the substantial path overlap issue in multimodal transport networks. Based on the generalized path-size formulation in *Eq.(2-5)*, they presented three types of path-size based on number of overlapped legs, travel time on overlapped legs, and distance on overlapped legs respectively. A leg was defined as a part of the path between two nodes in the network in which a single mode or service type is used. They estimated PSL Models based on their path-size definition and compared the results with MNL Model using a survey data

for train travelers in an urbanized corridor in the Netherlands. The results suggested that the path-size defined using the number of legs with γ=20 achieves substantially better results than the other definitions. However, their work did not focus on the mathematical background and even the exact path-size formulation under each path-size definition is not specified in their paper.

An extension of this work has specified the exact path-size formulations and proposed three path-size formulations for a route as function of path-size on sub-routes (Hoogendoorn-Lanser and Bovy, 2007). A route consists of home-end, sub-route, activity-end sub-route and train-part sub-route. Three path-size formulations were specified: 1) path-size formulation based on length L_{ir} of sub-route s_{ir} and the total number N_{na} of unique full routes using leg a in choice set C_n as:

$$PS_{irn} = \frac{1}{L_{ir}} \sum_{a \in \Gamma_{ir}} \left(\frac{l_a}{N_{na}} \right) \qquad \textit{Eq.(2-8)}$$

Where Γ_{ir} is the set of legs in sub-route s_{ir}; 2) PS formulation based on length of sub-route and the number of unique sub-routes and 3) PS formulation based on length of full route and the total number of unique full routes using leg. A MNL model and base PSL Models using the classic path-size formulation as in *Eq.(2-2)* and a set of PSL models using proposed PS formulations were estimated using survey data in an urbanized corridor in the Netherlands. The results demonstrated that PSL model with proposed PS formulations is superior to MNL and the base PSL, and travelers value overlap in the home end and the activity end of the trip negatively, whereas overlap in the train part is valued positively (Hoogendoorn-Lanser and Bovy, 2007). However, suggestion on

which proposed PS formulation to use is not given as they all achieves similar estimation results.

The advantages of PSL models are 1) they are able to address correlations between path alternatives, more specifically, path overlapping in the context of route choice modeling; 2) they are simple to estimate as the closed-form logit structure is maintained and therefore model estimation is relatively easier. The limitation is that they ignore correlations due to factors other than path overlapping.

2.5.3 Multinomial Probit

Daganzo and Sheffi (1977) introduced the MNP to model route choice decisions by assuming the error terms are normally distributed with a joint covariance matrix. The main advantage of MNP is that it allows arbitral specification of covariance structure. However, it does not have a closed form, and thus the model estimation requires heavy computing time. Besides, the covariance matrix becomes extremely complicated when there are large number of alternatives in the choice set, making it almost impossible to model route choice decisions in dense urban network. Therefore, it is only applied in simple network with constraint on number of alternatives, and hasn't been applied to large-scale network yet.

Yai et al. (1997) propose a MNP model with structured covariance matrix in the context of route choice in the Tokyo rail network. This work considerably limits the number of covariance parameters to be estimated. To further reduce the computational complexity, they have imposed maximum three alternatives for each OD pair in their application.

2.5.4 Generalized Extreme Value

The Generalized Extreme Value (GEV) is a family of models proposed by McFadden (1978) to account correlations among alternatives by assuming that the error terms of utility for all alternatives are jointly distributed as a generalized extreme value. When the correlation among alternative is zero, it becomes MNL. Contrary to the MNL model, the GEV models allow some correlations while maintaining closed form for selection probability.

Nested Logit Model (NL), as the best known relaxation of MNL, is obtained by partitioning the alternatives into different subsets or nests (Williams, 1977). The assumptions for NL are: 1) IIA only holds for alternatives within the same nest, and 2) one alternative cannot exist in multiple nests. The utility U_i for alternative i in nest B_k is:

$$U_i = V_k + V_{ki} + \varepsilon_k + \varepsilon_{ki} \qquad \text{Eq.(2-9)}$$

Where V_k is the deterministic utility of selecting nest B_k ; V_{ki} is the deterministic utility of selecting alternative i in nest B_k; ε_k is the error term of V_k and ε_{ki} is the error term of V_{ki}. The selection probability of alternative i is then:

$$P(i|C_n) = \frac{exp\left(\frac{V_i}{\lambda_k}\right)\left(\sum_{j\in B_k} exp\left(\frac{V_j}{\lambda_k}\right)\right)^{\lambda_k - 1}}{\sum_{l=1}^{K}\left(\sum_{j\in B_l} exp\left(\frac{V_j}{\lambda_l}\right)\right)^{\lambda_l}} \qquad \text{Eq.(2-10)}$$

The nesting parameter λ_i indicates the correlation between alternatives in nest B_k, and K is the total number of nests. When $\lambda_i = 0$ for all alternatives in choice

set C_n, it means there is no correlation between any alternatives , and the equation above becomes the same as MNL.

While NL assumes there is no overlapping between alternatives in different nests, the Cross Nested Logit (CNL) proposed by Vovsha (1997) relaxes this assumption. The selection probability of alternative i is given by the summation of the probability of choosing alternative i in each nest:

$$P(i|C_n) = \sum_{i \in B_k} P(k)P(i|k) \qquad \text{Eq.(2-11)}$$

Where $P(k)$ is the probability of choosing nest B_k and $P(i|k)$ is the conditional probability of choosing alternative i in nest B_k:

$$P(i|k) = \frac{(\alpha_{ik}\, exp(V_i))^{\frac{1}{\lambda_k}}}{\sum_{l=1}^{K}(\alpha_{il}\, exp(V_l))^{\frac{1}{\lambda_l}}} \qquad \text{Eq.(2-12)}$$

$$P(k) = \frac{\left(\sum_{i \in B_k}(\alpha_{ik}\, exp(V_i))^{\frac{1}{\lambda_k}}\right)^{\lambda_k}}{\sum_{l=1}^{K}\left((\alpha_{lk}\, exp(V_l))^{\frac{1}{\lambda_k}}\right)^{\lambda_k}} \qquad \text{Eq.(2-13)}$$

Where α_{ik} is the allocation parameter that indicates the portion of alternative i in nest B_k, under the constraint that $\sum_{k=1}^{K} \alpha_{ik} = 1$.

Prashker and Bekhor (1998) adopted the CNL to model route choice decisions where each link of the network corresponds to a nest, and each path to an alternative. The allocation parameter α_{ik} of path i in nest B_k with link k is then defined as a function with respect to the link length and the path length:

$$\alpha_{ik} = \frac{L_k}{L_i}\delta_{ik} \qquad \text{Eq.(2-14)}$$

Where L_k is the length of link k in nest B_k, L_i is the total length of path i, and δ_{ik} is the incident dummy indicating whether link k is along path i. This application of CNL on route choice model accommodates a rich correlation structure due to path overlapping. However, the nesting parameters are extremely hard to estimate due to large amount of nests. Wen and Koppelman (2001) proposed to estimate the nesting coefficient as a parameterized average of the allocation parameter:

$$\lambda_k = \left(1 - \frac{\sum_{i \in C_n} \alpha_{ik}}{\sum_{i \in C_n} \delta_{ik}}\right)^{\gamma} \qquad \text{Eq.(2-15)}$$

where γ is a parameter to be estimated. Based on this estimation approach, Ramming (2001) estimated CNL using route choice data collected on a small network in Boston and concluded that the PSL model with the generalized formulation outperforms the CNL model.

The Paired Combinatorial Logit model (PCL) is another GEV model that has been adopted for route choice application (Chu, 1989). It is further developed (Koppelman and Wen, 2000) and adapted (Prashker and Bekhor, 1998; Pravinvongvuth and Chen, 2005) to the route choice problem. Based on the CNL structure, they propose that each pair of path alternatives belongs to a nest, and the selection probability of path i is then:

$$P(i|C_n) = \sum_{j \in C_n, j \neq i} P(ij)P(i|ij) \qquad \text{Eq.(2-16)}$$

where $P(ij)$ is the probability of selection path pair (i,j) among all possible path pairs in choice set C_n and $P(i|ij)$ is the conditional probability of selecting path i from path pair (i,j), where:

$$P(i|ij) = \frac{exp(\frac{V_i}{1-\sigma_{ij}})}{exp(\frac{V_i}{1-\sigma_{ij}}) + exp(\frac{V_j}{1-\sigma_{ij}})} \quad \text{Eq.(2-17)}$$

$$P(ij) = \frac{(1-\sigma_{ij})\left(exp\left(\frac{V_i}{1-\sigma_{ij}}\right) + exp(\frac{V_i}{1-\sigma_{ij}})\right)^{1-\sigma_{ij}}}{\sum_{p=1}^{N-1}\sum_{q=p+1}^{N}(1-\sigma_{pq})\left(exp\left(\frac{V_p}{1-\sigma_{pq}}\right) + exp\left(\frac{V_q}{1-\sigma_{pq}}\right)\right)^{1-\sigma_{pq}}} \quad \text{Eq.(2-18)}$$

where σ_{ij} is the similarity coefficient between path i and path j, and N is the total number of alternative in the choice set. Prashker and Bekhor (1998) has defined the similarity coefficient σ_{ij} to be a function of path length L_i and L_j:

$$\sigma_{ij} = \left(\frac{L_{ij}}{L_i L_j}\right)^{\gamma} \quad \text{Eq.(2-19)}$$

Where L_{ij} is the overlapped length between path i and path j, and γ is a parameter to be estimated. Gliebe et. al. (1999) has modified this expression to:

$$\sigma_{ij} = \frac{L_{ij}}{L_i + L_j - L_{ij}} \quad \text{Eq.(2-20)}$$

The values of both similarity coefficients are constrained to be between 0 and 1, with 0 indicates there is no common link between path i and path j, while 1 indicates path i and path j are exactly the same. The same as CNL, parameters are very hard to estimate due to large amount of nests. For a choice set with N

alternatives, there are N(N-1)/2 nests in total. There hasn't been any application of CNL and PCL for route choice in large-scale network yet.

Nuzzolo et al. (2001) present a multi-dimensional logit-based route choice model for a schedule-based transit network. In this model, a boarding stop needs to be firstly selected, followed by path selection given the selected boarding stop. Both the stop selection and path selection are modeled with multinomial logit model. This model was reinforced later by adding departure time selection as an additional layer on top of boarding stop selection (Nuzzolo et al., 2012). The departure time selection is also modeled as a multinomial logit model. The advantage of this multi-dimensional structure is that it reveals the selection behavioural that passengers tend to select departure time based on all available information, and then select a nearby boarding stop given the departure time, and lastly select a path with given departure time and boarding stop. However, it does not address the path overlapping issue or the possible correlation between alternative paths under different boarding stops.

Hoogendoorn-Lanser and Bovy (2004) present a Hierarchical Nested Logit (HNL) choice model and a Multi-Nested Generalized Extreme Value (MN-GEV) choice model to address the correlation between alternatives for the home-end trips and activity-end trips with given choice set obtained from survey data for a multimodal network. In a HNL model, correlation between alternatives is accounted by clustering similar alternatives into nests. However, one alternative can only be allocated into one nest, cross nested alternatives are not allowed. In contrast to the HNL model, the MN-GEV model does allow differences in correlation along multiple choice dimensions. The model is

capable of accommodating the distinct dimensions symmetrically, thus overcoming the strict hierarchical structure of the multilevel HNL models.

2.5.5 Mixed Logit

Mixed logit model is becoming a popular mathematical structure for analyzing choice behaviour (Hensher and Greene, 2003; McFadden and Train, 2000). There are two main advantages of mixed logit over other discrete choice structures. Firstly it is able to incorporate unobserved heterogeneity across agents and choice situations through randomly distributed coefficients; and secondly, it relaxes the IID assumptions of the MNL model by addressing correlations between alternatives (Ben-Akiva et al., 1993; Hess et al., 2005). These two advantages correspond to the two versions of mixed logit model: 1) Random Coefficients Logit (RCL) specification and Error Components Logit (ECL) specification. RCL specification incorporates correlations across respondents through randomly distributed coefficients, while ECL specification considers the correlation over alternatives through an additive error structure, which is also referred as logit kernel. Therefore, RCL specification is more appropriate to address taste heterogeneity with panel data. Although these two versions provide different interpretations, they are proved to be mathematically equivalent (Train, 2009). This review will only focus on the RCL specification.

Revelt and Train (1998) proposed a framework to deal with data with repeated observations of respondents. The assumption is that user heterogeneity only exists among respondents and the same respondent making repeated choices are having the same tastes over choices. In their RCL specification, the utility of alternative i perceived by decision maker n with choice set C_n is specified as:

$$U_{in} = \boldsymbol{\beta_n X_{in}} + \varepsilon_{in} = \left(\overline{\boldsymbol{\beta_n}} + \boldsymbol{\eta_n}\right)\boldsymbol{X_{in}} + \varepsilon_{in} \qquad \text{Eq.(2-21)}$$

where X_{in} is the observed variables that relates to alternative i and decision maker n, ε_{in} is the unobserved error term that follows iid Gumbel distribution, and β_n is the randomized coefficients of these variables for decision maker n, with $\overline{\beta_n}$ being the deterministic taste preference and η_n is the randomized taste preference across decision makers. η_n can be specified to follow certain distributions, and the parameters of those distributions will be estimated from data. Commonly specified distributions in the literatures are: normal or lognormal distribution (Bastin et al., 2006; Ben-Akiva et al., 1993; Bhat and Gossen, 2004; Grigolon et al., 2014; Revelt and Train, 1998), truncated normal distributions (Revelt, 1999), triangular distribution (Revelt and Train, 2000; Train, 2001), uniform distributions (Revelt and Train, 2000; Train, 2001), and Rayleigh distributions (Siikamäki, 2001). Fosgerau and Bierlaire (2007) proposed a practical test based on semi-nonparametric techniques to examine whether a specific distribution is suitable.

Hess and Rose (2009) presented a framework to allow intra-person heterogeneity in mixed logit model and Hess and Train (2011) further specified to incorporate both inter-personal and intra-personal heterogeneity in mixed logit models. The utility of alternative i perceived by decision maker n at choice situation t is specified as:

$$U_{itn} = \boldsymbol{\beta_{nt} X_{in}} + \varepsilon_{int} = \left(\overline{\boldsymbol{\beta_{nt}}} + \boldsymbol{\eta_n} + \boldsymbol{\tau_{nt}}\right)\boldsymbol{X_{in}} + \varepsilon_{int} \qquad \text{Eq.(2-22)}$$

where β_{nt} is the randomized coefficients of these variables for decision maker n at choice situation t, ε_{int} is the unobserved error term that follows iid Gumbel

distribution, $\overline{\beta_{nt}}$ is the deterministic taste preference and η_n is the randomized taste preference across decision makers, and τ_{nt} represents the taste heterogeneity across choice decisions for person n.

Most mixed logit applications do not relate the observed attributes to the variation of taste, but only address the heterogeneity in general. This relationship can be modeled by relating the distribution parameters of β_n with observed attributes. Unfortunately, there are only few applications found in the literature that incorporate the observed attributes into the distribution of β_n with the mean of β_n to be a function of decision maker characteristics (Bhat, 1998, 2000). A recent work by Li et al. (2016) incorporates both observed and unobserved heterogeneity in route choice analysis by treating the random coefficients in three parts: agent specific term, OD pair specific term and choice situation specific term.

Few applications of adopts the RCL specification to deal driver's route choice decision with panel data (Bogers, 2009; Han et al., 2001), while Bekhor et al. (2002) use the ECL specification of mixed logit model to consider path overlapping problem.

There is only one application of Mixed Logit in modeling route choice in public transport networks, and it applied mixed logit to analyze passengers' transit route choice behaviour in Montreal (Eluru et al., 2012). The data used was obtained from a revealed preference survey conducted on university members. Passengers are assumed to be heterogeneous in travel time in train, total walking time squared and total number of transfers. Although not specified, they mentioned that they have interacted the total travel time with passengers' socio-

demographics attributes including age, gender and designation. However, the survey respondents are all from university including student, staff and faculty, which has introduced large sample bias into estimation results.

2.5.6 Other Modeling Frameworks

Except random utility framework, other frameworks have been used in the literature for modeling route choice behaviour. Models based on artificial neural networks (Yang et al., 1993), fuzzy logic (Henn, 2000; Rilett and Park, 2001), decision trees (Yamamoto et al., 2002) and other machine learning techniques (Chen et al., 2011; Nakayama and Kitamura, 2000; Park et al., 2007) have been proposed. This list of literature is not exhaustive but gives some existing alternatives to random utility framework.

Hurk et al. (2015) conducted a rule-based route choice model from smart card data. In this model, they identified the route selection rate based on six selection rules, namely, 1) first departure, 2) earliest arrival, 3) last arrival, 4) least transfer, 5) maximum route length, and 6) selected least transfer last arrival. The results indicate selected least transfer last arrival path achieves highest coverage most consistently. However, this approach has no qualitative indication on passenger's preferences over difference factors that affects path selection.

Chapter 3 – Modeling Stop-to-Stop Route Choice Behaviour from Smart Card Data

With the expanded usage of smart cards in public transport system, massive information on passengers' chosen routes becomes available with high accuracy. Researchers could benefit from this numerous route choice data in smart card by validating the choice sets and estimating route choice model parameters more accurately. In this chapter, a stop-to-stop route choice model using smart card data is presented to identify and quantify different aspects of travelling that affect passengers' stop-to-stop route choice decisions in the public transport network of Singapore. In the choice set generation stage, it evaluates six choice set generation approaches, namely, 1) labeling, 2) link elimination, 3) simulation, 4) k-shortest path, 5) nested labeling & link elimination, and 6) branch & bound approaches, to the multimodal public transport network in Singapore. A stop-to-stop route choice model is estimated based on final choice set with 100% coverage against one data set and the model is validated by examining the prediction performance against another two datasets using the estimated parameters.

3.1 Introduction

Along with the wide application of fare collection system using smart card in public transport network, massive amounts of passengers' actual route choices becomes available. Smart card data has been used in various ways in

transportation planning. There are three main categories of usage of smart card data to transit planners (Pelletier et al., 2011):

- Strategic-level. It includes studies that relate to network planning (Utsunomiya et al., 2006), user behaviour studies (Agard et al., 2006; Bagchi and White, 2005), and travel demand forecasting (Trepanier et al., 2009).
- Tactical-level. Studies at this level aim at optimizing service schedules (Utsunomiya et al., 2006), and deriving travel patterns (Agard et al., 2006; Hofmann et al., 2009; Jang, 2010),
- Operational level. At operational level, studies related to supply-and-demand indicators (Trépanier et al., 2009), as well as improvement in smart card system operations (Alfred Chu and Chapleau, 2008; Chapleau and Chu, 2007)

Researchers could also benefit from the numerous route choice data in smart card by validating the choice sets and estimating route choice model parameters more accurately. Despite there is no information on true origin-destination locations in the smart card data, the tremendous richness of passengers path selection from an initial boarding stop to the last alighting stop on public transport network serves as great data for investigating passengers' preferences from their stop-to-stop route choice decisions. The stop-to-stop route choice behaviour is useful in real-time traffic estimation simulators in which real-time smart card data serves as input to the simulator and the simulator simulates passenger route choice decision and outputs real-time traffic prediction. However, few studies use smart card data in route choice modeling. There still lacks literature validating the generated path choice sets using smart card data

and understanding how passengers make their route choice decisions in public transport network.

This chapter aims to identify and quantify different aspects of travelling that affect passengers' stop-to-stop route choice decisions in the public transport network of Singapore using smart card data. It implements six choice set generation approaches, namely, the 1) labeling, 2) link elimination, 3) simulation, 4) k-shortest path, 5) nested labeling & link elimination, and 6) branch & bound approaches, to the multimodal public transport network in Singapore, and comparatively evaluates these approaches in a comprehensive evaluation framework. Significant efforts have been made to reach 100% coverage for passenger journey observations in the final choice set with detailed investigation on identifying possible reasons for fail-to-generate paths. Route choice model is estimated based on final choice set with 100% coverage against one data set and the model is validated by examining the prediction performance against another two datasets using the estimated parameters.

The rest of the chapter is structured as follows. Section 3.2 presents the public transport network of Singapore examined in this thesis. In Section 3.3, the datasets used for evaluating choice set generation and route choice modeling are described. Section 3.4 describes the implementation and evaluation of the six choice set generation approaches for public transport network. Section 3.5 presents the stop-to-stop route choice model with model specification, estimation results, perception and evaluation, as well as prediction results.

3.2 Network Data

Google Transit data with stop information, service line information, service line frequencies and scheduled timetable for bus, MRT and LRT, was used to build up the network and compute link attributes such as in-vehicle travel time, walking time and waiting time.

To enable direct application of various choice set generation approaches on a public transport network, the network representation outlined by De Cea and Fernández (1993) is adopted. This involves creating "route segments" covering all stop-to-stop combinations within the same pattern as shown in Figure 3-1 (De Cea and Fernández, 1993).

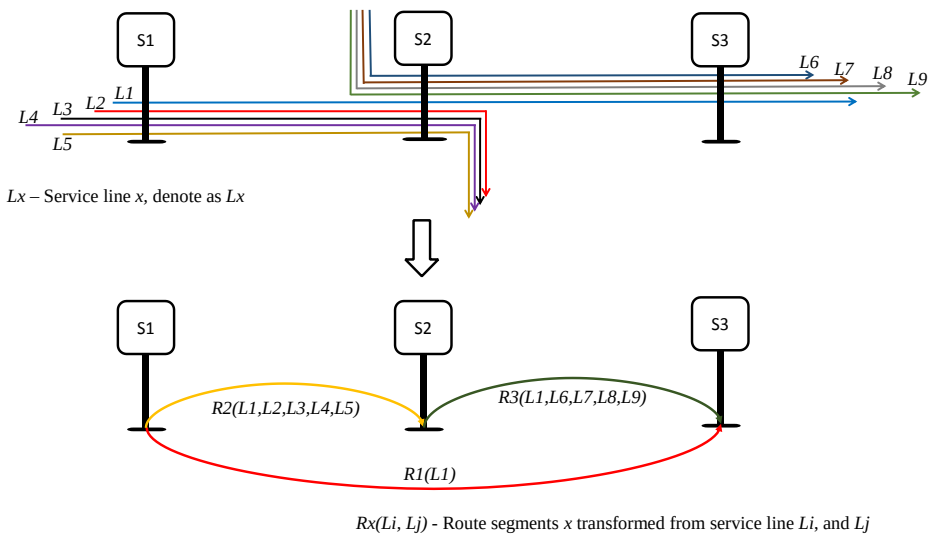


Figure 3-1 Demonstration of network representation

A route segment represents a direct connection between two stops served by at least one service line. All vehicle journeys that serve the route segment must run along the same sequence of roads. There are several advantages of this

network representation: 1) it addresses the common line issue in public transport network as all common lines serving the same stop pair are embedded in the same route segment and therefore, the effect due to common lines between stop pairs can be compensated in waiting time to embark on each route segment (De Cea and Fernández, 1993); 2) path overlapping issues can be easily addressed by computing the path-size with respect to overlapped route segment, which is more realistic as in public transportation networks, passengers perceive overlapping in terms of transfer points (Hoogendoorn-Lanser and Bovy, 2007). While in hyper-path approach, to the author's knowledge, no literature has been found on how to handle path-overlapping among hyper-path alternatives. 3) it helps to speed up the shortest path searching as route segments are created between stop pairs as long as they can be reached via at least one service line without a transfer, regardless of their actual distance. In this network reformulation, the computational time for each shortest path search is proportional to the number of transfers along the path regardless of the actual length or traversed stops of that path. Empirical test also demonstrates the shortest path searching time (0.07sec) on the proposed network are generally 10 times faster than shortest path searching time (0.69sec) on the super-Network presentation with separate links for different service lines and transfer between lines (Carlier et al., 2003); and 4) Aggregation of common service lines on route segment gives freedom to model the boarding behaviour and interactions between passengers and service vehicles in an agent-based simulation environment. In an agent-based supply simulator, an agent is given a set of attractive lines, which is the set of common lines on route segment, at each

transfer point derived from the selected path in the route choice model. Various boarding mechanisms such as first-come-first-board, board with seat only, and etc., could be easily applied.

The disadvantage of this approach is the number of route segments created increases quadratically with the number of stops along each service line. This requires larger memories for data storage. However, given the fast advancement of computational power and data storage, it deems not to be a practical issue.

Each stop/station in the public transport network is treated as a node in the network while links are route segments with direct connection between two stops served by at least one service line without transfer. Walking links were then added on to the network between all stop pairs which are less than 500 meters away from each other, assuming a Euclidian distance. The total network contains 4,718 nodes, 447,432 transit route segments, and 54,486 walking links.

In-vehicle travel time is retrieved from schedule data by selecting the scheduled travel time among all service lines on the route segment. Travel time on walking links was estimated by assuming a constant walking speed of 4km/hour for every passenger, with reference to the recommended pedestrian walking speed for most conditions by The Federal Highway Administration (Rouphail et al., 1998). Waiting time can be calculated in several ways. If departure time of a passenger is given, waiting time can be computed by selecting the minimal waiting time with respect to scheduled bus arrival time. Here the departure time is the arrival time at the boarding stop. If departure time is not available, expected waiting time can be calculated by estimating service line frequency

with an assumed passenger arrival distribution. In this work, expected waiting time is based on the overall frequency of common service lines on the route segment as defined in Eq. (3-1):

$$t_{wait,R_x} = \frac{1}{2\sum_{L_i \in R_x} f_i} \qquad \textit{Eq. (3-1)}$$

Where t_{wait,R_x} denotes the expected waiting time to embark on route segment R_x; L_i is a service line traversing on the route segment; and f_i is the frequency of line L_i.

Note that the network configuration is specified to facilitate path choice set generation only. Route choice model estimation depends on the definition of paths which is generic in this context and could be applied to other cities directly. In this chapter, a path is defined as a sequence of transit legs (route segments) with different transport services that passenger takes from the initial boarding stop to the final alighting stop in the public transport network while the travel mode is defined as a transport service type such as Bus, rapid transit (including MRT, and LRT), or walk.

3.3 Route Choice Data

EZLink card is the contactless smart card used in the public transport network of Singapore. It is the dominant fare payment approach in Singapore's public transit system (EZlink, 2013). Passengers have to both tap in whilst boarding a bus and tap out when alighting from a bus, while for MRT and LRT, passengers have to tap their EZlink card when entering and exiting MRT and LRT stations.

Around 4 million of passenger journeys on Singapore's public transport system are recorded in EZlink card daily (PublicTransport@sg, 2013).

A passenger journey in EZLink card is defined as a series of single-mode trips with maximum 5 transfers within 2 hours. Each transfer can take up to 45 minutes. Previous work on analyzing public transport journeys in London using smart card data has suggested 45 minutes as optimal transfer thresholds for bus-to-bus journey, and 30 minutes for the bus-to-metro and metro-to-metro transfer thresholds (Seaborn et al., 2009). This work has adopted such threshold with additional time-distance constraints to better classify the ownership of a transit leg to a particular passenger journey. The additional constraints are:

- The transfer distance for each transfer should be less than 1km.
- The total transfer time should be less than 1.2 times the sum of expected walking time and maximum waiting time.

These two constraints are added by analyzing heavily detoured records in smart card data with manual investigation by plotting them into geographic interface system. These two constraints are added to filter out error journey records for 1) early tapping out of EZLink card, and 2) combined two journeys into one journey by taking advantage of the EZlink recording strategy. A recent paper by Robinson et al. (2014) describes the impact of imposing such constraints on improve the quality of smart card data.

The major limitation of smart card is that it does not include any information on actual origin and destination. Due to this data limitation, all choice set generation approaches in this work are applied to generate stop-to-stop path

choice set only, where the first stop denotes the initial boarding stop of the passenger journey while the second stop denotes the final alighting stop of the passenger journeys.

Three data sets are sampled from all the observed stop-to-stop pairs in smart card over the whole Singapore on 11 April 2011. As the smart card data does not record transfer stations between rapid transit services, it is not possible to estimate route choice decisions for pure MRT/LRT trips using smart card data. Therefore, when sampling stop-to-stop pairs, if the sampled stop-to-stop pair is a pure MRT/LRT trip, it is discarded and resampled until a stop-to-stop pair with either boarding stop or alighting stop is a bus stop is drawn. Each data set contains 1,000 stop-to-stop pairs.

An interesting finding from smart card data is the high path dominancy for stop-to-stop paths. Vast majority (around 82.8%) of stop-to-stop pairs have single observed path selection and there are less than 0.1% stop-to-stop pairs which are observed to have more than 1 dominant paths in the smart card data. Here a path is identified as dominant if no less than 5% of total passengers on that stop pair were observed to select the path and the total observed passengers on the stop-to-stop pair should be more than 50. Such path dominancy issue is also observed on the bus network in Japan (Schmöcker et al., 2013).

3.4 Choice Set Generation

Choice set generation is an important topic in large-scale transportation networks. It has computational advantages as enumeration of alternative is not practical in large networks, particularly if route choice has to be determined in

real time, such as with journey planner systems. Most importantly, the set of path alternatives from which passengers chose their paths helps us to better understand their path selection behaviour. Besides, with a pre-generated choice set, path overlapping issues can be explicitly handled via various theoretical correction approaches (Bekhor et al., 2006; Prato, 2009). An adequate choice set should satisfy certain basic requirements: 1) path alternatives should be logical. For example, path alternatives should not contain heavy detours or loops; 2) path alternatives should be feasible in terms of time, space, service availability, and physical availability; 3) path alternatives should likely be chosen by a passenger; and 4) ideal choice set should contain at least path alternatives which have been observed or recorded. The composition and quality of generated path choice set also strongly influences subsequent parameter estimation in route choice model and route choice prediction (Bekhor et al., 2006; Hoogendoorn-Lanser, 2005; Prato and Bekhor, 2006; Van der Waerden et al., 2004).

3.4.1 Implementation of Methods

In this section, six choice set generation approaches are introduced and implemented with special treatments to make them suitable for generating paths alternatives on a multimodal public transport network. For transfers on paths, only a single walking transfer between stops or a direct transfer at stop is allowed. Consecutive walks along a path is not considered as feasible. Particularly for stop-to-stop choice set, additional path feasibility check is imposed on all approaches to ensure that for each stop-to-stop path, the first link

should not be a walking link. Note that this specific constraint only applies to stop-to-stop choice set.

3.4.1.1 Labeling Approach

The labeling approach generated path sets by searching the least cost paths using different cost functions, called labels (Ben-Akiva et al., 1984). It addresses passengers' heterogeneity on perceiving the least cost path between origin and destination. The labeling approach has not been explicitly applied on the public transport network in the literature. This work proposes the 10 labels tabulated in Table 3-1, to be used on the multimodal public transport network.

Table 3-1 Description of implemented labels in the labeling approach

Labels	Label Description
1	Minimal total in-vehicle travel time
2	Minimal number of transfers
3	Minimal walking transfer time
4	Maximized travel on MRT
5	Maximized travel on Bus
6	Minimal waiting time at transfers
7	Minimal total travel time 1 (in-vehicle +waiting + transfer walking)
8	Minimal total travel time 2 (in-vehilce+5*waiting+3*transfer walking)
9	Minimal total travel time 3 (in-vehilce+5*waiting+5*transfer walking)
10	Minimal total travel time 4 (in-vehilce+5*waiting+8*transfer walking)

Label 1 to 6 defines different single path attributes as "least cost" function in path searching. For example, label 2 - Minimal number of transfers, will

generate an alternative path with minimal number of transfer between initial boarding stop and final alighting stop. Label 7-9 weights multiple path attributes differently to achieve different "least cost" definitions in the path searching. In label 7, the least cost is defined to search for alternative path of which passengers perceive the in-vehicle travel time, the waiting time and walking time equally, while in label 8-10, the waiting time and walking time are both more negatively weighted than the in-vehicle time; and the waiting time is perceived more, equal and less negatively than the walking time respectively. Note that, due to the non-additive fare scheme in Singapore, minimal monetary cost is not examined in labeling approach.

3.4.1.2 Link Elimination Approach

The link elimination approach adopted in this work iteratively searches shortest path/paths by removing the links along the searched paths one by one. When each link along the first shortest path has been eliminated once, all the links along this path will be eliminated, and the iteration will move to the next generated path. The stop criterion is defined as the total number of generated paths, which is set as 30 in this work. Cost consists of in-vehicle travel time, waiting time, walking time, and transfer penalty.

3.4.1.3 K-shortest Paths Approach

The K-shortest paths approach of Yen (1971) is adopted in this thesis, with the following modifications to make it suitable for multimodal public transport network: 1) instead of comparing node sequences, it has been modified to compare link sequences for determining links to eliminate, as parallel links exist in our public transport network; 2) all walking links at origin are eliminated to

avoid generating infeasible paths. K is set to 30 in this work as the maximum size of observed paths is only 15 from the smart card data, which is described in Section 3.2.

3.4.1.4 Simulation Approach

In the simulation approach, in-vehicle travel time, walking time, and waiting time for each link are all randomly sampled from an independent and identically distributed normal distribution with mean equals to its original value and standard deviation set to 5 times of the original value. To avoid drawing a negative value, the absolute value is taken. 50 draws of randomized travel times were performed for each sample OD. The selection of sampling distribution and number of draws takes into consideration of the maximum size of observed paths, coverage, and computational time.

3.4.1.5 Branch & Bound Approach

The branch & bound approach can be adapted to the multimodal network by treating each possible transfer as a branch. This thesis implements branch & bound approach based on the work by Hoogendoorn-Lanser et al., (2007). Additionally, special treatments to path logicality, common lines and preference on transfers are added. Table 3-2 depicts the constraints implemented for branch & bound approach with newly added constraints highlighted in bold. There are two new logical constraints on walking leg added in the branch & bound approach. Constraint 2 is added because the choice set generation approaches are used to generated stop-to-stop path choice set. Therefore, the first leg and the last leg for a stop-to-stop path should be transit legs. The maximum allowable number of transfer for branch & bound approach is set to be five. This

is because computing a choice set for one stop-to-stop pair becomes practically infeasible (more than five hours) for certain stop-to-stop pairs when maximum allowed number of transfer is larger than four. This constraint was not deemed an issue, since more than 99.95% of the observed paths in the smart card data have four or less transfers. Note that this type of transfer does not include transfers on rapid transit network, as such transfer is not recorded in smart card data.

Table 3-2 Description of implemented constraints for the branch & bound approach

Constraints	Type	Constraint Description
1	Logical	Path is loopless
2	**Logical**	**First leg or last leg is not a walking leg**
3	**Logical**	**No Consecutive walking legs**
4	Feasible	Service Availability, such as there must be a service line goes to destination stop.
5	Feasible	Access availability, such as whether a stop can be walked from another stop
6	Feasible	Maximum allowable number of transfer (<=5)
7	Behavioural	Bounds on trip attributes: waiting time, walking time, in-vehicle time, maximal allowable number of transfers
8	Behavioural	Spatial constraints-preference on travelling in the direction of destination
9	**Behavioural**	**No selection of extra transfers when there are adequate common lines (>=3)**
10	**Behavioural**	**In selection of transfer stops with more common lines than stops with fewer common lines**
11	**Behavioural**	**When there is direct service available, no selection of more than 1 transfer.**
12	**Behavioural**	**No selection of path with two more transfers than least transfer path**

3.4.1.6 Nested Labeling & Link Elimination Approach

The nested labeling and link elimination approach combined the searching heuristics of the labeling approach and link elimination approaches. It performs a link elimination inside each labels as defined in Table 3-1. For label 2-Minimal number of transfers, it will acquire a link elimination approach to find 10 paths under the cost definition, while for all other labels, only three paths are to be searched by the link elimination approach in each label. This configuration was inspired by the significant impact of number of transfers on passengers' path selection. It was calibrated to achieve good coverage against the observations in smart card data.

3.4.2 Evaluation of Methods

This section proposes a comprehensive evaluation framework to qualitatively and quantitatively analyze choice set generation approaches and presents validation and evaluation results on choice set generation approaches based on actual route choice data collected by smart card in Singapore. The proposed evaluation framework can be divided into four parts: computational performance, coverage tests, and composition evaluation of choice sets, and analysis of fail-to-generate paths. All approaches were implemented in R using the "igraph" library for network building and shortest path searching, the "hash" library for building hash tables, and the "data.table" library for data storage. All tests were run on an Ubuntu system with Intel® Core™ i7-3770 CPU @ 3.40GHz ×8 with 16G RAM. Only one CPU core was used.

3.4.2.1 Computational Performance

Table 3-3 tabulates the mean of standard deviation of the computation time per stop-to-stop pair for each choice set generation approach.

Table 3-3 Mean and standard deviation of computational time for generating choice set for 1 stop-to-stop pair

choice set generation Approach	Mean (sec)	Standard Deviation (sec)
Labeling Approach	13.9	2.9
Link Elimination (N=30)	25.6	4.5
Simulation Approach (50 draws)	45.7	9.2
K-Shortest Path (K=30)	99.5	13.6
Nested Labeling & Link Elimination	28.4	6.5
B&B Approach (max no. of transfer = 5)	-	-

The result for the B&B approach is not available on examined data set as it could not finish generating 200 choice sets over two weeks. It was forced to be terminated. As mentioned in the data set section, some stop-to-stop pairs in the data set can only be reached by at least four transfers. The computational time for the branch & bound approach is expected to increase exponentially with respect to the total number of transfers along the paths. An evaluation of the branch & bound approach on another dataset which contains fairly amount of paths with direct service lines has shown that it has an average running time of 108 seconds with standard deviation of 1020 seconds. It demonstrates that the branch & bound approach is not practically applicable to dense public transport networks. The branch & bound approach is therefore not included in the rest of

the evaluations. All the other five choice set generation approaches are able to generate path alternatives in reasonable time. The labeling approach takes the least computational time on average while the k-shortest path approach takes the longest computational time.

3.4.2.2 Coverage Test

With full-scale records on almost all passengers' actual path selection from smart card data, it becomes possible to evaluate the choice set coverage in new dimensions. Three coverage evaluation indexes, *Passenger Journey Coverage, Efficient Coverage*, and *Passenger Path Coverage* are proposed as follows:

- **Passenger Journey Coverage** is defined as the percentage of passengers whose chosen alternative is available in the generated choice set. It takes passengers' path selection into consideration. It represents the percentage of passengers who will choose their paths from the generated choice set by each approach and indicates the effectiveness of choice set generation approach.
- **Efficient Coverage** is defined as the percentage of generated alternatives being a recorded alternative. It indicates how efficient the choice set generation approaches are in producing observed paths and not producing unobserved paths.
- **Passenger Path Coverage** is defined as the percentage of observed paths being available in the generated choice set. It implies the comprehensiveness of alternatives in choice set generated by choice set generation approach.

The difference between passenger journey coverage and passenger path coverage is that the former is a weighted version of the latter, by the number of passengers observed on each path. A choice set with high passenger path

coverage but low passenger journey coverage indicates the choice set misses attractive paths and it will introduce biases in model estimation. Efforts should be put to enhance the choice set generation to include the attractive paths. If the generated choice set has high passenger journal coverage but low passenger path coverage, it also indicates that the investigations should be carried on to the fail-to-generated paths as they might be due to network error or path record error.

Efficient coverage indicates how efficient the choice set generation approaches are, in terms of producing observed paths. For a fixed number of observed paths per stop-to-stop pair, efficient coverage reaches highest if the choice set generation approach only produce paths that are observed, while it decreases along with generation of unobserved paths.

Table 3-4 Passenger journey coverage, efficient coverage and passenger path coverage of choice set generation approaches

Item	choice set generation approach	Passenger Journey Coverage	Efficient Coverage	Passenger Path Coverage
1	**Labeling**	96.78%	68.27%	61.11%
2	**Link elimination (N=30)**	98.16%	5.72%	84.66%
3	**Simulation (50 draws)**	95.74%	5.16%	60.93%
4	**K-shortest path (K=30)**	98.35%	5.39%	84.72%
5	**Nested labeling & link elimination**	98.65%	6.24%	84.94%
6	**Combined all above approaches**	98.73%	3.13%	87.11%

Passenger journey coverage is the most important as it affects the subsequent route choice model estimation mostly. Table 3-4 tabulates the coverage

evaluation indexes computed based on generated choice sets for data set 1 using different choice set generation approaches. In the last row, it presents the coverages for combined choice set in which it consolidates all paths generated by the labeling, link elimination, simulation, k-shortest path, and nested labeling & link elimination approaches. In the subsequent sections, we denote it as combined approach. All the approaches achieves high passenger journey coverage more than 95%.

Out of the five single choice set generation approaches, the labeling approach achieves highest efficient coverage over 50%, almost 10 times higher than other approaches. It indicates the labeling approach is very efficient in terms of generating path that are mostly selected by passengers. Detailed investigation on the uncovered passenger journeys shows most of these paths are competitive paths, which is not able to be captured by any label. These paths have similar costs in each label but the selection of different transfer stops varies. This finding is the inspiration to the nested labeling & link elimination approach. In the nested labeling & link elimination approach, performing link elimination inside each label helps to identify paths with similar cost but different transfer stops. The result is promising as the nested labeling & link elimination approach becomes the most effective single choice set generation approach in producing favorable paths as it achieves highest passenger journey coverage and highest passenger path coverage among all single choice set generation approaches. The performance of the k-shortest path and link elimination approaches are most

comparable. They both achieve high passenger journey coverage and passenger path coverage, but low efficient coverage.

Despite the simulation approach has a higher number of draws (S=50) than the K value in k-shortest path approach (K=30) and the N value in link elimination approach (N=30), its performance is the poorest. As indicated by its lowest efficient coverage and passenger path coverage, it generates more paths than other approaches with lowest coverages on observed paths. This is because the simulation approach generates substantial amount of non-relevant paths in different draws. Besides, the generated path set by the simulation approach is not completely reproducible due to the nature of its stochasticity. To investigate the variation of generated paths using the simulation approach, 200 draws from a normal distribution were undertaken 100 times on a stop-to-stop pair with 11 observed alternatives. The average size of the generated choice sets is 150 with little fluctuation.

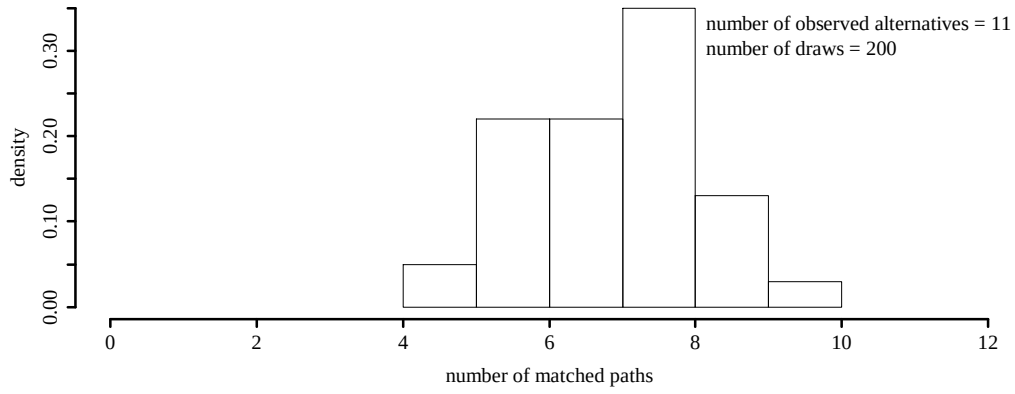


Figure 3-2 Density distribution of number of matched paths using simulation approach

As shown in Figure 3-2, 35% of the time, it is able to find 8 observed alternatives whereas for more than 50% of choice sets, the matched alternatives are fewer than 8. Therefore, in terms of actual implementation, it is not suggested to only employ stochastic choice set generation approaches. Instead, if used, stochastic choice set generation approach should always be combined with one deterministic choice set generation approach to guarantee that a substantial amount of observed paths will be generated.

The combined approach reaches higher passenger journey coverage and passenger path coverage than any single choice set generation approach. High passenger journal coverage indicates that the path alternatives the combined approach failed to generate are not frequently selected by passengers. This high passenger journey coverage also supports the proposed network presentation which aggregates the common lines into route segments and represents paths only in terms of transfer stops. Conventional paths with specified service line can be easily obtained, while maintaining the same coverage, by disaggregating common lines on transit legs along each path in the choice set. The computational time to achieve the same coverage will increase exponentially if common lines are not aggregated into route segments for path searching.

3.4.2.3 Path Composition in Choice Sets

A thorough evaluation of choice set generation approaches does not only consider the coverage, but also the quality of generated path sets. Composition evaluation helps to understand various qualities of generated choice sets, to infer network characteristics, and to evaluate the feasibility of choice sets for

subsequent route choice model estimation. The following four criteria are proposed to evaluate the quality of generated choice sets by each approach:

- Size of generated path choice set
- Diversity of path alternatives in the path choice set
- Additional Number of transfers with respect to least transfer path
- Variations of path attributes in path choice set

With observed paths from smart card data, size of generated path choice set serves as a good indicator on whether this choice set is not suitable for route choice modeling. A choice set generation approach that always generate fewer alternatives than the observed paths might result very inaccurate prediction in route choice model estimation as it will exclude a lot of alternative in estimation. A choice set generation approach that generates extremely large size of alternatives will also impose computational inefficiency and potential modeling difficulty in route choice model estimation.

Figure 3-3 presents density for the size of generated choice sets for all six approaches and compares them with the observed path choice sets. All approaches have explicit constraints on the maximum number of paths, out of which, the labeling and simulation approaches generate path alternatives that are much fewer than the number of labels and the specified number of draws. The size of the generated choice set for the rest single approaches are around their upper limits respectively. The labeling approach generates an extremely small size of choice sets which is always fewer than the observed path

alternatives in data set. It is therefore not suitable to adopt labeling method alone for route choice modeling.

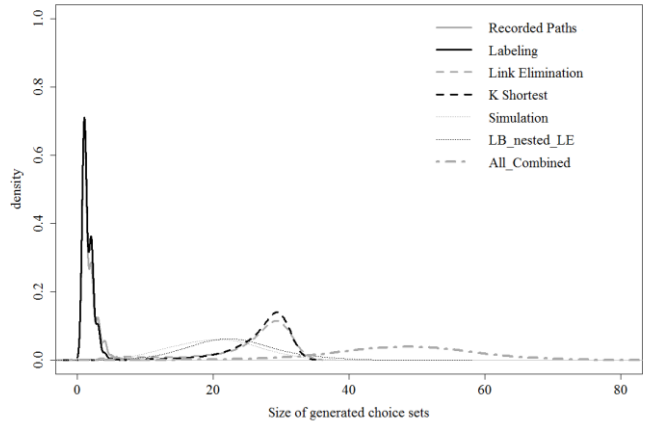


Figure 3-3 Density of size of generated choice sets

Diversity of path alternatives in path choice set helps to infer network condition and evaluate the efficiency of choice set generation approach in generating diverse paths. To analytically investigate the diversity of path alternatives in the path choice sets, the concept of path-size which is a well-recognized correction term to address path overlapping among alternatives in path-size logit models was employed in this work (Ben-Akiva and Ramming, 1998).With the introduction of route segments as a non-transferred link between stops via a single mode in the public transport network, the following path-size definition is proposed to better cope with the perceived overlapping in the public transport network:

$$PS_{in} = \sum_{r \in \Gamma_i} \left(\frac{t_r}{T_i}\right) \frac{1}{\sum_{j \in C_n} \delta_{rj}} \qquad \textit{Eq. (3-2)}$$

Where PS_{in} is the path size of path i for Stop-to-Stop pair n; r is the indexed route segment, Γ_i is the set of all route segments along path i; t_r is the travel time on route segment r; T_i is the total travel time on path i; C_n denotes the set of path alternatives for Stop-to-Stop pair n; δ_{rj}equals to 1 if route segment $r$is on path j and 0 otherwise; T_i denotes the total travel time of path i. Under this definition, a unique path will have a path-size equal to one while for overlapped paths, the path-size is smaller than 1.

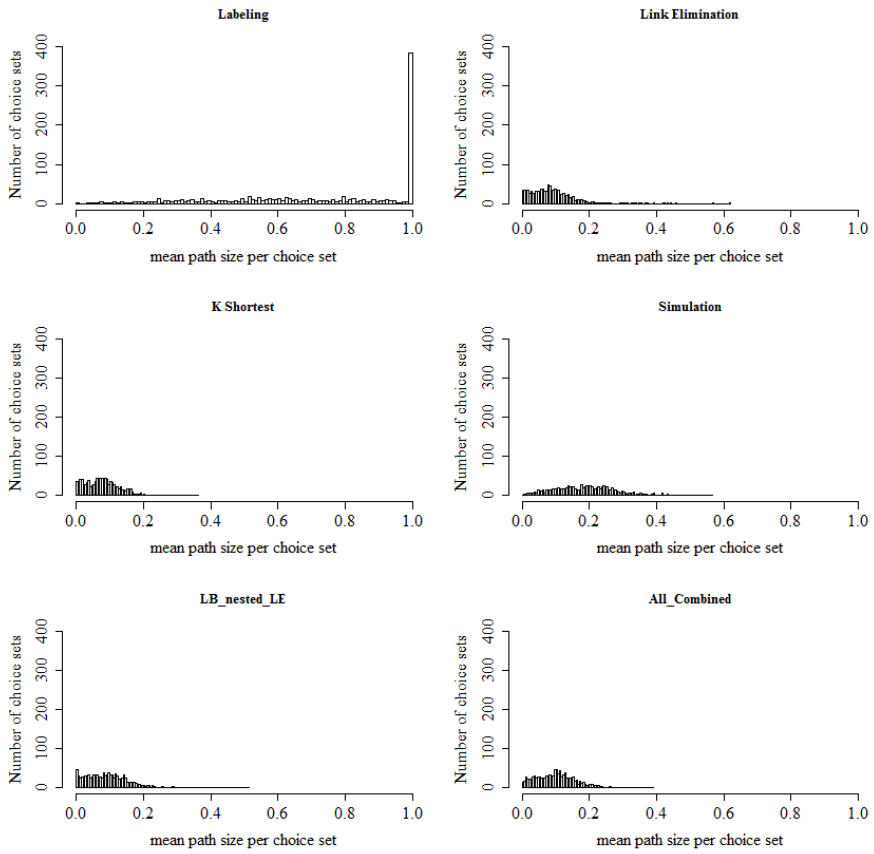


Figure 3-4 Histogram of mean path-size value of generated choice sets

The histogram of mean path-size value for choice sets generated by each approach is shown in Figure 3-4. The k-shortest paths approach generates the

most heavily overlapped paths in choice set as almost all the stop-to-stop pairs have mean path-size smaller than 0.4 while the labeling approach is producing a lot of diverse path in the choice sets with path-size value equal to 1.0. In general, the k-shortest path, link elimination, and nested labeling & link elimination approaches generate much more similar paths than labeling an simulation approach. This is expected as the searching heuristics in these approaches are based on eliminating one or certain links on the searched path. Similar paths are likely to be generated by these approaches. The reason for the extremely lower overlapping by the labeling approach is partially due to it is always trying to search path with different cost definitions, yet heavily due to its extreme small size of the generated choice sets. The low mean value of path size indicates there exists heavy degree of path overlapping in choice set and therefore, for such choice sets, correlation due to path overlapping must be addressed in route choice modeling.

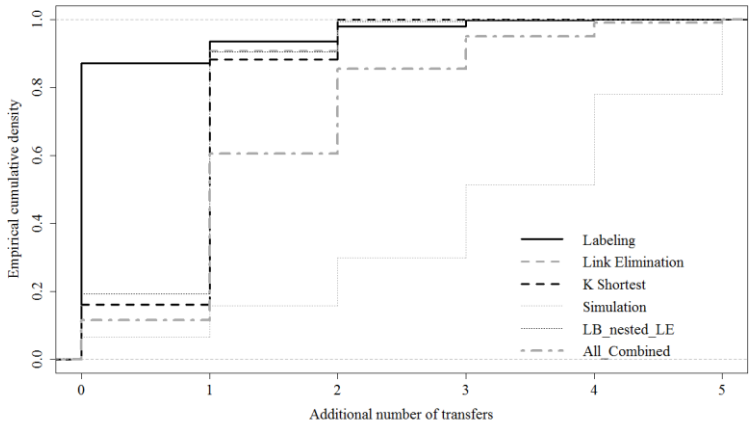


Figure 3-5 ECDF of additional number of transfers with respect to least transfer path

Additional number of transfer with respect to the path with the minimal number of transfer in the choice set is a good indicator of public transport path composition. By analyzing the EZlink card data, more than 80% of the passengers select paths with minimal number of transfers while passengers are not observed to select paths with 2 more transfers than the path with minimal transfer. A choice set generation approach that generates a large amount of paths with much more additional transfers is not considered good in producing relevant paths. Figure 3-5 shows the ECDF of additional number of transfers with respect to minimal number of transfers in each choice set. In consideration to the high passenger journey coverage of the nested labeling & link elimination approaches, it is most effective single choice set generation approach as it producing much fewer irrelevant alternatives than other approaches with high coverage.

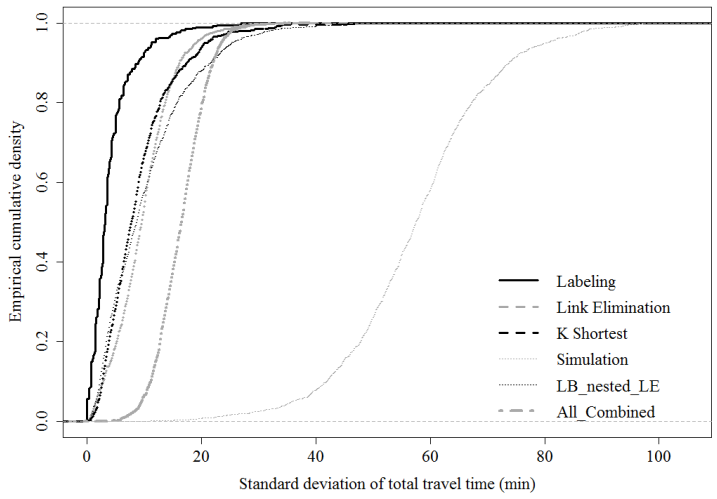


Figure 3-6 ECDF of standard deviation of total path travel time per stop-to-stop pair

The ECDF of the standard deviation of total travel time including waiting time, in-vehicle travel time and walking time is depicted in Figure 3-6. Given the limited number of path alternatives generated by the labeling approach, the standard deviation of total path travel time does not statistically meaningful for the labeling approach. The k-shortest paths, link elimination and nested labeling & link elimination approaches generate paths with less variation in travel time while the simulation approach produces more deviated paths in terms of total travel time. The results are expected as the other three approaches are all searching new shortest paths based on certain link elimination heuristics, it is likely to generate paths with only small deviations at the eliminated link.

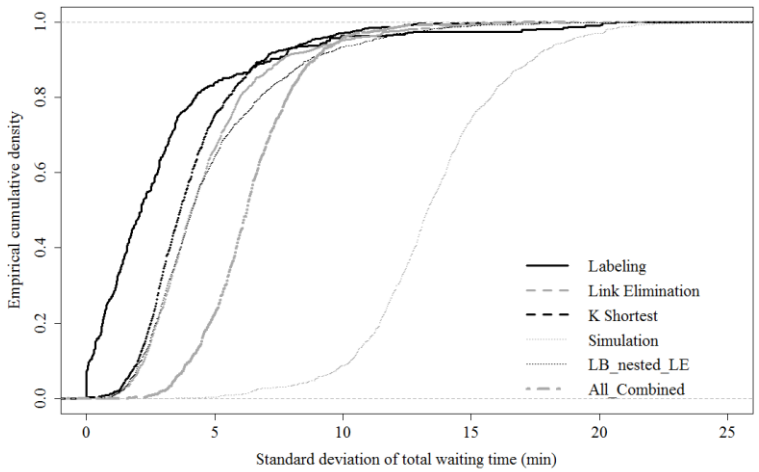


Figure 3-7 ECDF of standard deviation of total waiting time per stop-to-stop pair

The waiting time is one key attribute in path alternatives for the public transport network. It also serves as a good indicator on network performance. As shown in Figure 3-7, paths generated by the k-shortest path, link elimination and nested labeling & link elimination approaches have smaller variation in waiting time

while the simulation approach produces paths with relatively more variant waiting times. It follows the same pattern as total transit time. More than 80% of the paths in the combined choice set are not heavily deviated to the average waiting time, as the 80-th percentile for all ECDFs are smaller than 10 minutes. It indicates that the network has a good connectivity such that there are plenty of path alternatives with reasonable waiting time.

3.4.2.4 Analysis of Fail-to-generate Paths

For the recorded 62,302 passenger journeys in data set 1, there were around two hundreds recorded paths with 792 passenger journeys that are failed to be generated by in the combined choice set. Denoting these paths as "Fail-to-generate" paths, further investigation has been carried out to identify possible reasons. Three major categories of failed cases were identified and they are:

- Algorithm deficiency
- Network data issues
- Abnormal recorded paths

Algorithm deficiency exists mainly in the labeling approach. The major drawback of the labeling approach is that it is not able to recognize distinct path alterative with similar costs. It is always able to recognize the most dominant path in a choice set, but fails to capture less dominant paths. It is because majority of the most dominant path is the least cost path for several labels in the examined network. The labeling approach is also not able to capture paths with less waiting time at certain transfer stop but longer total waiting time. Although "least waiting time" is one label in the labeling approach, it is searching paths

with the least total expected waiting time. While in reality, the passenger might choose certain routes simply because they have least waiting time at that decision point. The searching heuristics in the link elimination, k-shortest path, and simulation approach are in favor of identifying less dominant paths neither. Although eventually, it would enumerate all path alternatives if stop criteria are relaxed to infinite, it is not practically.

Network data issues consist of 1) missing stops along service lines, 2) missing service lines, 3) Erroneous scheduled travel time, 4) inaccurate walking time as walking time is estimated based on directly geometric distance, not representing actual walking time, and 5) Subjective specification of transfer penalty.

There are five types of abnormal recorded paths found in the fail-to-generate paths:

1) Path with additional transfer when the passenger is already on the most direct service line - which might be due to accident or intentional stop for certain purposes such as: grocery shopping, buying lunch or newspaper, and etc.;
2) Paths contain loops;
3) Path with transfer between stops which are more than 10 km away from each other – which is probably due to taking a taxi for part of the trip to void being late; and
4) Path with getting off bus service line one stop earlier and walking much longer distance to transfer.

3.4.3 Final Choice Set

The final choice sets used for route choice modeling combines all paths generated the labeling, link elimination with N=30, k-shortest path with k=30, and simulation approach with number of draws = 50 approaches, as well as nested link elimination and labeling approaches. As explained in Section 3.4.2.4, detailed investigations have been carried out to identify possible reasons for the observed paths that are failed to be generated. Efforts were made to increase the passenger journey coverage by correcting the network errors and eliminate abnormal paths in observations. The final choice set reaches 100% coverage after correcting network errors and removing abnormal paths.

However, due to limitation of the smart card data, it is not possible to evaluate the complete OD path choice set as actual origins and destinations are not observed in smart card data. A feasible method to generate a complete origin-to-destination path choice set based on the stop-to-stop path choice set generated by the choice set generation approaches could be:

- Step 1: Generation of initial boarding stop choice set for each origin which includes all the stops within prefix radius distance from the origin.
- Step 2: Generation of final alighting stop choice set for each destination which includes all the stops within 1km radius distance from the destination.
- Step 3: Form stop pair choice set for each origin-destination pair by enumerating all the possible combination of initial boarding stop and final alighting stop.

- Step 4: Generation of stop-to-stop path choice set for each stop pair using choice set generation approaches mentioned in section 3.2.
- Step 5: Concatenation of stop pair choice and stop-to-stop path choices to form origin-to-destination choice set for each origin and destination pair.

This method could be easily implemented given the origins and destinations. If the passengers are assumed to only select initial boarding stop/final alighting stop which is less than 1Km always from the origin/destination, then the stop-to-stop path choice set generated using choice set generation approaches will cover the same observed paths as the complete OD path choice set generated using method above. This is a reasonable and practical assumption. Therefore, in this case, applying the choice set generation approaches only to stop-to-stop path choice set is sufficient and more appropriate to evaluate the performance of these choice set generation approaches.

3.5 Modelling Stop-to-Stop Route Choice Behavoiur

In this work, we seek to identify and quantify the different aspects of travelling that affect passengers' route choice decisions on Singapore's public transport network. A random utility model is considered in this work. It is assumed that each passenger n chooses a route i with maximum possible utility level U_{in} among a set of available alternatives C_n. The utility U_{in} of path i faced by passenger n is defined as:

$$U_{in} = V_{in} + \varepsilon_{in} = \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}_{\boldsymbol{in}} + \varepsilon_{in} \qquad \textit{Eq. (3-3)}$$

Where V_{in} is the observed utility for path $i$in choice situation *n* that is assumed to have a linear relationship with path attributes $\boldsymbol{X_{in}}$; $\boldsymbol{\beta}$ is a vector of coefficients that represents passenger preference on path attributes and ε_{in} denotes the unobserved error components which follow i.i.d. Gumbel distribution. This model is a multinomial logit model. Denote it as MNL, the probability that path i will be chosen by passenger *n* given the MNL model is then:

$$P(i|C_n) = \frac{e^{\boldsymbol{\beta^T X_{in}}}}{\sum_{j \in C_n} e^{\boldsymbol{\beta^T X_{jn}}}} \qquad \textit{Eq. (3-4)}$$

To address the correlation between the alternative routes due to path overlapping, we consider Path-Size Logit (PSL) model with path size defined as in Eq. (3-2). In this PSL model, the logarithm of path-size is added to each path utility. In the PSL model, the probability of choosing path i by passenger n becomes:

$$P(i|C_n) = \frac{e^{\boldsymbol{\beta^T X_{in}} + \beta_{PS} ln(PS_{in})}}{\sum_{j \in C_n} e^{\boldsymbol{\beta^T X_{jn}} + \beta_{PS} ln(PS_{jn})}} \qquad \textit{Eq. (3-5)}$$

Where β_{PS} is the coefficient to estimate against the logarithm of path-size value.

Both MNL and PSL models are estimated with regard to the final choice set for dataset 1.

3.5.1 Model Specification

In-vehicle travel time, walking time, waiting time, and number of transfers are the three key path attributes that are commonly considered in public transport

route choice models. Fares are commonly included in route choice models if the fare is not collected at flat rate. It is essential when analyzing value of travel time. All of these attributes are expected to have negative impact on path utility in route choice models. Apart from these conventional attributes, other miscellaneous aspects such as weather (Sumalee et al., 2011), and topological directness of paths (Raveau et al., 2011) are also considered in the utility function by some modelers as well.

The path attributes $\boldsymbol{X_{in}}$ used in these models include the total number of transfers along the path, the total path cost, three time components and one mode specific constants. It applies to both MNL model and PSL model. This work considers three time components: the in-vehicle travel time, the waiting time at each boarding stop/station, and the walking time during transfer. All three time components are in unit of minutes. Note that access walk time and egress walk time are obviated in this work as only stop-to-stop path selection is considered in the public transport network. One mode specific constant indicate whether a path utilizes rapid transit services is also included. Mode constant indicating that a path is conducted on rapid transit only is not feasible in this work as smart card data does not record down transfers on rapid transit network and all these observations were exclude in the data set. Constant indicate whether a path utilizes Bus is therefore the only complement of the mode-specific constant for rapid transit and it is normalized to have zero coefficient.

3.5.2 Estimation Results

Estimation results are tabulated in Table 3-5. All the estimated parameter coefficients are statistically significant at 95% confidence interval with expected right signs for both models.

Table 3-5 Estimation results of route choice model

Variables	MNL			PSL		
	Coefficient	**Std.Err.**	**t-stats**	**Coefficient**	**Std.Err.**	**t-stats**
total in-vehicle time (min)	-0.202	0.005	-37.6	-0.233	0.006	-37.18
total waiting time (min)	-0.382	0.010	-37.9	-0.380	0.010	-38.62
total walking time (min)	-0.836	0.012	-67.1	-0.822	0.012	-66.8
number of transfers	-5.76	0.040	-145	-5.79	0.040	-143.9
total cost (S$)	-5.62	0.464	-12.1	-4.20	0.471	-8.91
Dummy constant if rapid transit is used	6.60	0.609	10.8	6.08	0.611	9.96
path-size	NA	NA	NA	0.444	0.040	11.17
Summary of Statistics:						
Number of observations	61996			61996		
LL(0)	-238149.9			-238149.9		
LL(β)	-9101.5			-9049.4		
Likelihood Ratio Test	458096.9			458201.1		
rho-squared	0.9618			0.9620		
adjusted rho-squared	0.9618			0.9620		

The likelihood ratio test statistics for null hypothesis that all the coefficients are zero, is -2(LL(0)-LL(β)), where LL(0) is the log-likelihood value of the model when all the coefficients are zero while LL(β) stands for the log-likelihood value when all the coefficients are the estimated value at convergence. It is used to test the null hypothesis that all the coefficients are zero. It follows chi-square

distribution with K degree of freedom where K is the number of parameters to estimate. The high rho-squared value justifies this route choice model is able to capture passengers' actual path selection very well.

As both models are performed over data set 1 for the same 1000 sampled stop-to-stop pairs and cover the same passenger journeys, both the final log-likelihood and the adjusted ρ^2 serve as useful and direct measurements on model estimation accuracy. The likelihood ratio test statistics for the null hypothesis of generic attributes for MNL model and PSL model is much larger than the critical value of 7.879 at 99.5% confidence interval. It proves that the PSL model is superior to the base MNL model in terms of estimation accuracy. The higher adjusted ρ^2 of PSL model supports this claim.

3.5.3 Perceptions and Valuations

Out of all traditional explanatory parameters, namely, in-vehicle travel time, waiting time, and walking time, walking time is having the most negative coefficient, which implies that passengers have the strongest disfavor on longer walking time. It might be partially because the comfort level of walking in Singapore is low due to the humid weather. The total in-vehicle travel time is having the least negative coefficient. This is partially due to the big magnitude of in-vehicle travel time, but it also implies that passengers do not sensitively perceive the differences in in-vehicle travel time as comparing to other path attributes. The estimated coefficient on number of transfers is the most negative over all choice sets implying that number of transfers has significant impact on passenger's route selection.

The value of time for time component t can be obtained as follows:

$$VOT_t = \frac{\beta_t}{\beta_{cost}} \qquad \textit{Eq. (3-6)}$$

where, t stands the time component of interest, such as time in-vehicle time, the walking time or the waiting time; β_t is the estimated coefficient of time component t, β_{cost} is the estimated coefficient on travel cost. Time value is round 3S$ per hour for total in-vehicle travel time, 6S$/per hour for total waiting time, and 11S$/hour for total walking time. The walking time is more valued than waiting time. It indicates that passengers in Singapore are more willing to wait than to walk. It might be because of the hot and humid tropic weather in Singapore such that the walking environment is not as comfortable as waiting at bus stop or train station. It's worth noting that the data used in this study is from November. It has the most rain days in Singapore. Dislike to walk might be largely due to this seasonal effect as well.

As for number of transfers, the model predicts around 24 minutes of in-vehicle travel time per transfer. This marginal rate of substitution for transfer is higher than the value obtained in the literature for other public transport networks (Eluru et al., 2012; Raveau et al., 2011). One reason is that obviating the access and egress walk reduce the number of transfers overall. It is also because the examined network is much larger with much better connectivity than other networks in the literature with around 90% of the trips conducted on this network have no more than two transfers.

The estimate of mode specified constant for rapid transit has a large positive coefficient. It reveals the behavioural preference on traveling on MRT or LRT instead of pure buses when both types of paths are available.

3.5.4 Prediction Results

To examine the prediction performance of the estimated route choice model, another two data sets: data set 2 and data set 3, are used to test the prediction performance for the estimated model. The choice set for each data set is generated using the same procedure as described in Section 3.4. Table 3-6 below presents the prediction results on different data sets based on behavioural parameters estimated in the route choice model. As shown from the table, both models achieve high prediction performance as indicated by the high adjusted rho-squared values. The PSL model always outperforms the MNL model in predication accuracy for both data sets by achieving the highest LL(β) and adjusted ρ^2.

Table 3-6 Prediction results on different data sets using estimated parameters

	Data Set 2		**Data Set 3**	
Summary of Statistics:	**MNL**	**PSL**	**MNL**	**PSL**
Number of observations	61445	61445	58565	58565
LL(0)	-228327.8	-228327.8	-216342.2	-216342.2
LL(β)	-9037.2	-8883.6	-9226.5	-8929.7
Likelihood Ratio Test	438581.2	438888.3	414231.5	414825.0
rho-squared	0.9604	0.9611	0.9574	0.9587
adjusted rho-squared	0.9604	0.9611	0.9573	0.9587

3.6 Summary

This chapter presents a stop-to-stop route choice model using smart card data to identify and quantify different aspects of travelling that affect passengers' stop-to-stop route choice decisions in the public transport network of Singapore. It evaluates six choice set generation approaches in the multimodal public transport network of Singapore. The estimate stop-to-stop route choice behaviour is validated by examining the prediction performance against another two datasets using the estimated parameters.

The contributions of this chapter are threefold. First, it implements and evaluates different choice set generation approaches qualitatively and quantitatively. Special treatments on six conventional choice set generation approaches were specified to make them suitable for generating choice sets in public transport network, and the nested labeling & link elimination is proposed with inspiration of the observed passenger path selection behaviours. To the authors' knowledge, it might be the first to undertake experiments to examine variation of generated paths by the simulation approach and to apply the labeling approach with specified labels, and the nested labeling & link elimination approach in public transport network. The evaluation of choice set generation approaches conducted in this work is summarized in Table 3-7 with comparison to the two reviewed literature in section 2.4.3.

Second, significant efforts have been made to reach 100% coverage for passenger journey observations in the final choice set with detailed investigation

on identifying fail-to-generate paths. Third, a stop-to-stop route choice model is estimated using large amount of actual route choice data collected by smart card on a large-scale urban network in an international city. Passengers' actual route choice behaviours have been interpreted from the estimated model. Prediction accuracy is also examined based on another different datasets.

Table 3-7 Summary of the evaluations with comparison to studies in the literature

Reference	**(Bekhor et al., 2006)**	**(Rieser-Schüssler et al., 2012)**	**This work**
Network	highway network in Boston, U.S.	Road Network in Zurich, Switzerland	Public Transport Network in Singapore
Nodes	13000	408636	4718
Links	34000	882120	501918
Data Type	Survey at University	GPS data	Smart card data
Total Observations	188 observations	36,000 observations with 2,434OD	3,670,339 observations with 687,199 Stop-to-Stop pairs.
Evaluated choice set generation approaches	Labeling Link Elimination K Shortest Path Simulation	Link Elimination K Shortest Path Simulation B&B	Labeling Link Elimination K Shortest Path Simulation Nested Labeling & Link Elimination B&B Approach
Common Evaluated Criteria	Path Coverage Computation Time Size of Choice Set Number of Links	Path Coverage Computation Time Path Similarity Path Distance	Path Coverage Passenger Coverage Efficient Coverage Computation Time Analysis of fail-to-generated paths No of Transfers Path Total Travel Time Path Waiting Time
Route Choice Model Estimation	Estimation of Route Choice Models	N.A	Estimation of Route Choice Models
Path Selection Prediction	N.A	N.A	Prediction Performance on two sampled data sets from the whole observations

Chapter 4 - Addressing Path Overlapping for Public Transport Network

The Path-Size Logit (PSL) aims at capturing correlations among path alternatives by including a correction term -path size, to the Multinomial Logit (MNL). Several path size formulations have been proposed in the literature and applied to various route choice models extensively in road networks. However, the path-size formulation has rarely been studied in the public transport context considering the special characteristics of public transport network. This chapter proposes a new path-size formulation for route choice modeling in public transport networks. The new path size formulation accounts not only the correlation due to overlapping of path alternatives along traversing roads, but also the correlation due to overlapping boarding station. The new path size formulation was comparatively analyzed against a MNL and a base PSL with direct application of existing path size formation, in route choice model estimation and prediction using smart card data, on the public transport network in Singapore. The results show that the new path size formulation has outperformed other models in both estimation and prediction.

4.1 Introduction

Discrete choice analysis under the random utility theory frame work is the most widely applied modeling tool for route choice modeling. It assumes that an agent has a randomized unobserved utility and the decision is always made to maximize this utility. Under this framework, Multinomial Logit (MNL) model

is the most commonly used discrete choice model in practice due to its simplicity. However, MNL models assume that the unobserved utilities for different alternatives are identically and independently distributed. This assumption is not valid in the context of route choice, particularly due to paths overlapping (Bovy and Stern, 1990). Efforts have therefore been made to overcome this restriction by making a deterministic correction of the utility for overlapping paths. One easy and practical way is to add a deterministic correction term to the path utility function while maintaining the simple path selection structure of MNL models, such as C-Logit (Cascetta et al., 1996) and Path-Size Logit (PSL) (Ben-Akiva and Bierlaire, 1999). These models intend to heuristically approximate path correlation by adding a correction term to the utility function subject to the degree of path overlapping. They have computational advantages by retaining the simple Logit structure and yet outperform MNL models.

However, the existing path-size formulation has rarely been studied in the public transport context considering the special characteristics of public transport network. This chapter addresses the path overlapping in public transport network considering the unique characteristics of public transport services such as frequencies of service lines and transfers. New definition of path size value is proposed considering the boarding service frequencies at overlapped transfer stops with its detailed derivation. Furthermore it is analyzed using actual route choice data collected through smart card on the public transport network in Singapore. To the author's knowledge, this is the first time that frequency is considered in the path size formulation.

Public transport network in Singapore is used in this work. The general information of this network is described in section 3.2. Majority of the bus lines in Singapore are served less than 10 minutes per bus in peak hours and both MRT and LRT are served less than 5 minutes per train throughout the day (SMRT, 2014a, b; Transit, 2014). Besides, there is a high degree of common lines at important transfer stops. Given the high frequent and intensive services in Singapore's public transport network, the new path size formulation and subsequent route choice analysis is conducted on a frequency-based network. The smart card in Singapore records both passengers boarding and alighting. It therefore provides adequate passenger path selection data for route choice modeling. Stop-to-stop route choice models are applied to examine the proposed path size formulation for stop-to-stop path selection. This work utilizes the same datasets and generated choice sets as specified in Chapter 3

The chapter is structured as follows. In section 4.2, formulation for path-size on public transport network is then proposed with detailed derivation. Then a set of route choice models including a PSL model with proposed path-size formulation are estimated for comparison and applied for prediction in section 4.3. In the end, Section 4.4 concludes the chapter.

4.2 Methodology

In this section, the classic path-size formulation and the new frequency-based path-size formulation are derived with explicit definition of assumptions and properties in the context of public transport network. This derivation is based on the theory of aggregated alternative (Ben-Akiva and Lerman, 1985). Similarly to road networks, the definition of path overlap is based on the

overlapped links in the public transport networks. Particularly for public transport network, each link stands for a road segment from one bus stop or train station to its subsequent stop/station along the same service line.

4.2.1 Derivation of Path-Size Formulation

Given the choice set C_n for an OD pair n, let l denotes a path alternative in the choice set, and r denotes link that belongs to l. According to the theory of aggregate alternatives (Ben-Akiva and Lerman, 1985), the aggregate alternative – link r then has a utility defined as

$$U_{rn} = \max_{l \in C_n}(V_{ln}^r + \varepsilon_{ln}^r) \qquad \text{Eq. (4-1)}$$

Where V_{ln}^r ires the part of deterministic utility of elementary alternative l that contributes to the aggregated alternative r; $\varepsilon_{ln}^r$is assumed to be independently and identically distributed (IID) Gumbel distribution with μ as positive scale parameter and η as the mode of the distribution. If aggregated alternative is a combination of the whole elementary alternatives, V_{ln}^r is then equal to V_{ln}. In route choice context, aggregated alternative – link r is only a combination of partial elementary alternatives.

According to the property of Gumbel distribution, if $(x_1, x_2, \dots, x_J)$ are J independent Gumbel distributed random variables with parameter (η_1, μ), (η_2, μ) ,…, (η_J, μ) respectively, then $\max(x_1, x_2, \dots, x_J)$ is also Gumbel distributed with parameters:

$$\left(\frac{1}{\mu} ln \sum_{j=1}^{J} e^{\mu \eta_j}, \mu\right)$$

Therefore, this utility U_{rn} is also a Gumbel distribution with parameters:

$$\left(\frac{1}{\mu} ln \sum_{j=1}^{J} e^{\mu V_{ln}^{r}}, \mu\right)$$

The utility of aggregated alternative – link r can also be written as:

$$\begin{aligned} U_{rn} &= E\left(\max_{l \in C_n}(V_{ln}^{r} + \varepsilon_{ln}^{r})\right) + \varepsilon_{rn} \\ &= \frac{1}{\mu} ln \sum_{l \in C_n} \delta_{rl} e^{\mu V_{ln}^{r}} + \varepsilon_{rn} \end{aligned} \qquad \textit{Eq. (4-2)}$$

Where ε_{rn} is iid Gumble distributed with $(0, \mu)$ and δ_{rj} is the link – path incident value. It equals to 1 if link *r* is along path *i*, and it is 0 otherwise.

Define the average deterministic utility $\overline{V_{rn}}$ of the paths using link r as:

$$\overline{V_{rn}} = \frac{1}{\sum_{l \in C_n} \delta_{rl}} \sum_{l \in C_n} \delta_{rl} V_{ln}^{r} \qquad \textit{Eq. (4-3)}$$

Under the assumption that all elementary alternatives are iid Gumbel distributed with the same deterministic utility for all $l \in C_n$, it can be proved that $\overline{V_{rn}} = V_{ln}^{r}$, and therefore:

$$U_{rn} = \overline{V_{rn}} + \frac{1}{u} ln \sum_{l \in C_n} \delta_{rl} + \varepsilon_{rn} \qquad \textit{Eq. (4-4)}$$

That is, overlapping of elementary alternatives has introduced a positive correction to the size of an aggregated alternative. Accordingly, overlapping of aggregated alternative should result a negative correction of the utility to an elementary alternative (Frejinger, 2008). Therefore, the size correction for an elemental alternative - path l due to overlapping of aggregated alternative – link r can then be defined as:

$$\frac{1}{u} ln \frac{1}{\sum_{l \in C_n} \delta_{rl}}$$

For road networks, the size of a path is assumed to be proportional to the length of the overlapped links among path alternatives in the choice set, whereas for public transport networks, as passengers value travel time much more than travel distance and given the same travel distance, the travel time varies on different public transport services, definition based on travel time on the overlapped links might be more appropriate. Therefore, in this case, the contribution to the path-size of a path from each link is assumed to be proportional to the travel time on that link. If t_r denotes the travel time along link *r* and T_i is the total travel time over all links along path *i*, then the path-size formulation on public transport network can be derived as:

$$PS_{in} = \sum_{r \in \Gamma_i} \left(\frac{t_r}{T_i}\right) ln\left(\frac{1}{\sum_{j \in C_n} \delta_{rj}}\right) \qquad \textit{Eq. (4-5)}$$

Denote this formulation as the direct adoption of path-size formulation to public transport networks. This formulation is similar to the path-size correction factor in *Eq.(2-5)*. Note that this formulation can be applied to both frequency-based networks and schedule-based networks.

4.2.2 Additional Path-Size Formulation

In public transport network, path overlapping does not only apply to overlapping of links, but also overlapping of boarding stations. At each boarding station, passengers have the freedom to keep current path or select another path. Unlike road network, the selection of another path is not only depending on remaining path, but also affected by the waiting time at each boarding station.

Let's follow the same derivation in previous section according to the theory of aggregate alternatives by treating boarding station *s* as aggregated alternative

as well. It can be easily prove that the utility U_{sn} for aggregated alternative s of elementary alternative l is:

$$U_{sn} = \overline{V_{sn}} + \frac{1}{u} ln \sum_{l \in C_n} \delta_{sl} + \varepsilon_{sn} \qquad \textit{Eq. (4-6)}$$

Under the assumption that all elementary alternatives are IID Gumbel distributed with the same deterministic utility for all $s \in C_n$, it can be proved that $\overline{V_{sn}} = V_{ln}^{s}$, and therefore, Therefore, the size correction to an elemental alternative - path l due to overlapping at boarding stops with other elementary alternatives can then be defined as:

$$\frac{1}{u} ln \frac{1}{\sum_{l \in C_n} \delta_{sl}}$$

In a frequency-based network, the overlapping of boarding station can therefore be assumed to be proportional to the frequency on the subsequent transit leg. Under this assumption, a new path-size formulation is proposed as an additional factor particularly to address the correlation due to overlapped boarding stops/stations for frequency-based public transport network:

$$PS_{in}^{node} = \sum_{s \in S_i, s \neq origin} ln \left(\frac{f_{si}}{\sum_{j \in C_n} \delta_{sj} f_{sj}} \right) \qquad \textit{Eq. (4-7)}$$

Where S_i denotes the collection of all boarding stops along path i; f_{si} denotes the boarding frequency over all common service lines at boarding stop s along path i, and δ_{sj} equals one if stop s is also a boarding stop along path j, otherwise it is zero. Note that the overlapping of origin should not be included into this path-size formulation as it is compulsory to start from origin and

therefore the aggregated utility function in *Eq. (4-1)* for origin does not hold. In the context of this work, the origin means the initial boarding stop.

It implies that path containing overlapped boarding stations with less frequent services lines should have a smaller path-size and therefore more negative deduction in the total path utility. While a path containing overlapped boarding stations with more frequent services will have a larger path-size and less deduction in the total path utility.

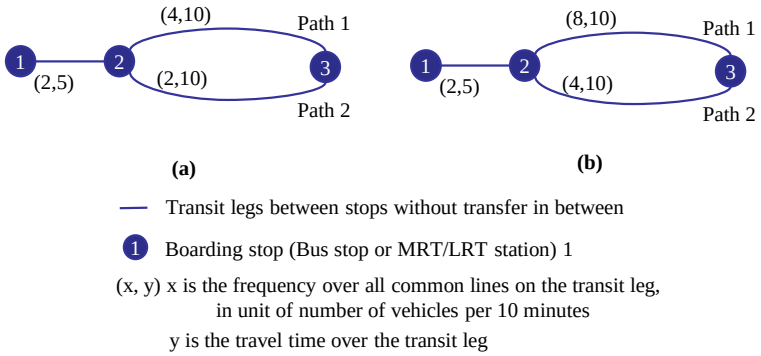


Figure 4-1 Example of path correlation due to common boarding stop

Two examples of path overlapping in public transport network are illustrated in Figure 4-1. In both (a) and (b), the initial boarding stop is at node 1, and the destination is at node 3. Path 1 and path 2 overlapped with each other on the transit leg from node 1 to 2 with common boarding stop at node 2. Travel time from node 1 to node 2 is assumed to be 5 minutes, while travel time from node 2 to node 3 on both path 1 and path 2 is 10 minutes. In (a), the boarding frequency is 4 vehicle per 10 minutes along path 1 and 2 vehicles per 10 minutes along path 2, while in (b) the boarding frequency is 8 vehicle per 10 minutes along path 1 and 4 vehicles per 10 minutes along path 2. Assuming all other

deterministic utilities are the same for both paths, it is more likely that passenger will select to board public transport services along path 1 at node 2 because it has higher frequency but the relative selection ratio between path 1 and path 2 should be the same for both scenarios. The expected waiting time, base PS value (from Eq. (4-5)) and proposed additional path-size value (from Eq. (4-7)) are computed for the hypothesized choice scenario shown in Figure 4-1 and the results are tabulated in Table 4-1.

Table 4-1 Comparison of path size values in different formulations

		Expected waiting time	Base path-size value	Additional path-size value
Figure (a)	**path 1**	1.25 minutes	-0.231	-0.405
	path 2	2.5 minutes	-0.231	-1.099
Figure (b)	**path 1**	2.5 minutes	-0.231	-0.405
	path 2	5 minutes	-0.231	-1.099

In both scenarios, the base path-size value captures the similarity of path 1 and path 2 due to path overlapping while the additional path-size value captures the correlation between path 1 and path 2 due to overlapping at node 2, and it indicates that the additional new path-size formulation is able to capture passengers' preference on selecting the transit leg along path 1 from node 2 to node 3 because it has higher frequency. In both scenarios, the relative selection ratio between path 1 and path 2 should be the same for both scenarios. However, the difference in expected waiting time are not the same for both scenarios as there is 1.25 minutes waiting time difference in scenario (a) and there is 2.5 minutes time difference in scenario (b). In this case, introducing the additional

path-size value helps to correct the non-linearity of waiting time among paths sharing the same boarding stops.

4.2.3 Path Size Logit Model

In this section, the PSL model with the proposed path-size formulation as in Eq. (4-7) is specified. The utility of path i in choice situation n - U_{in} is:

$$\begin{aligned} U_{in} &= V_{in} + \beta_{PS} PS_{in} + \beta_{PS}^{node} PS_{in}^{node} + \varepsilon_{in} \\ &= \boldsymbol{\beta^T X_{in}} + \beta_{PS} PS_{in} + \beta_{PS\_node} PS_{in}^{node} + \varepsilon_{in} \end{aligned} \qquad \textit{Eq. (4-8)}$$

Where V_{in} is the observed utility for path i in choice situation *n* that is assumed to have a linear relationship with path attributes $\boldsymbol{X_{in}}$; $\boldsymbol{\beta}$ is a vector of coefficients that represents passenger preference on path attributes; PS_{in} is the base path-size value as defined in Eq. (4-5) with coefficient β_{PS} to be estimated, PS_{in}^{node} is the proposed additional path-size value as defined in Eq. (4-7) with coefficient β_{PS}^{node} to be estimated, and ε_{in} denotes the unobserved error components which follow i.i.d. Gumbel distribution.

The probability that path i will be chosen in choice situation *n* given the PSL model is then:

$$P(i|C_n) = \frac{e^{\boldsymbol{\beta^T X_{in}} + \beta_{PS} PS_{in} + \beta_{PS}^{node} PS_{in}^{node}}}{\sum_{j \in C_n} e^{\boldsymbol{\beta^T X_{jn}} + \beta_{PS} PS_{jn} + \beta_{PS}^{node} PS_{jn}^{node}}} \qquad \textit{Eq. (4-9)}$$

4.3 Results and Analysis

Based on the data collected on actual route choices in the Singapore's Public Transport network by smart card, a MNL model, a base PSL model with direct

application of classic path-size formulation as defined in Eq. (4-5) and a PSL model with new path-size formulation proposed in this chapter are estimated.

The utility for path i in the MNL model is:

$$U_{in} = \boldsymbol{\beta^T X_{in}} + \varepsilon_{in} \qquad \textit{Eq. (4-10)}$$

The utility for path i in the base PSL model is:

$$U_{in} = \boldsymbol{\beta^T X_{in}} + \beta_{PS} PS_{in} + \varepsilon_{in} \qquad \textit{Eq. (4-11)}$$

While the utility for path i in the new PSL model is:

$$U_{in} = \boldsymbol{\beta^T X_{in}} + \beta_{PS} PS_{in} + \beta_{PS\_node} PS_{in}^{node} + \varepsilon_{in} \qquad \textit{Eq. (4-12)}$$

The same datasets as in Chapter 4 are used for model estimation and analysis. The description of data sets sampled from smart card data can be found in section 3.3. Dataset 1 is used for model estimation, and dataset 2 and 3 are used for prediction. The estimation results are presented in Table 4-2.

As all models are performed over data set 1 for the same 1000 sampled stop-to-stop pairs and cover the same passenger journeys, both the final log-likelihood and the adjusted rho-squared serve as useful and direct measurements on model estimation accuracy. The likelihood ratio test statistics for the null hypothesis of generic attributes for MNL model and both PSL models are much larger than the critical value of 7.879 at 99.5% confidence interval. It proves that both PSL models are superior to the base MNL model in terms of estimation accuracy. The new PSL model with additional proposed frequency-based path-size formulation further delivers good improvement over the base PSL model, by

achieving 30 more points in final log-likelihood and 0.003 increase in the adjusted rho-squared.

Table 4-2 Estimation results of MNL, base PSL and PSL with new PS formulation

Variables	MNL		base-PSL		new-PSL	
	Coefficient	t-stats	Coefficient	t-stats	Coefficient	t-stats
total in-vehicle time (min)	-0.202	-37.6	-0.233	-37.18	-0.228	-36.14
total waiting time (min)	-0.382	-37.9	-0.382	-38.62	-0.378	-37.75
total walking time (min)	-0.836	-67.1	-0.822	-66.8	-0.820	-66.25
number of transfers	-5.76	-145	-5.79	-143.9	-6.04	-113.1
total cost (S$)	-5.62	-12.1	-4.20	-8.91	-4.16	-8.76
Dummy constant if rapid transit is used	6.6	10.8	6.08	9.96	5.85	9.56
Base path-size	NA	NA	0.444	11.17	0.437	11.27
New path-size	NA	NA	NA	NA	-0.268	-7.75
Summary of Statistics:						
Number of observations	61996		61996		61996	
LL(0)	-238149.9		-238149.9		-238149.9	
LL(β)	-9101.5		-9049.4		-9019.358	
Likelihood Ratio Test	458096.9		458201.1		458261.1	
rho-squared	0.9618		0.9620		0.9621	
adjusted rho-squared	0.9618		0.9619		0.9621	

The estimates for path attributes other than path-size are similar in all models. Note that the estimated coefficient for the new path-size is negative, indicating a preference over paths with shared boarding stops. One possible reason could be path with more shared boarding stops with other paths might be more reliable in case of congestion or accident, as passengers have options to switch to other paths. The new path-size is related with frequency and therefore may correlate

with total waiting time. Empirically, the path-size and the total waiting time have a low correlation in the data set, and these two attributes are totally different transformations of frequency.

Coincidence ratios (CR) for each model are also computed to measure how well the modeled path selection distribution overlaps with the observed path selection distribution. The formula to compute CR is analogue to the conventional CR formulation used for validating trip distribution methods as in (Beagan et al., 2007):

$$CR = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\sum_{i\in C_n} min(obs_i, est_i)}{\sum_{i\in C_n} max(obs_i, est_i)}\right) \qquad \textit{Eq. (4-13)}$$

where n is the index for a stop-to-stop pair with N denotes the total number of stop-to-stop pairs used for model estimation; obs_i stands for the observed probability of selecting path i, and est_i denotes the estimated probability of selecting path i in choice set C_n. The CR values calculated are 0.943, 0.948, and 0.951 for the MNL model, the base-PSL model and the new-PSL model respectively. It indicates the path selection distribution produced by the PSL with new PS model is more close to the observed path selection distribution.

To examine the prediction performance of the proposed PS formulation, these three models estimated based on data set 1, were used to predict the passenger path selection in data set 2 and data set 3. The prediction results obtained from each route choice model are listed in Table 4-3. As shown from the table, the PSL model with new PS formulation outperforms the rest in predication accuracy by achieving the highest final log-likelihood LL(β) and adjusted rho-squared in both data sets.

Table 4-3 Prediction results of MNL model, base PSL model and new PSL model

		MNL	base-PSL	new-PSL
Data Set 2	Number of observations	61445	61445	61445
	Number of estimates	6	7	8
	LL(0)	-228331.3	-228331.3	-228331.3
	LL(β)	-8787.6	-8631.4	-8576.8
	Likelihood Ratio Test	439087.2	439399.7	439509.0
	rho-squared	0.9615	0.9622	0.9624
	adjusted rho-squared	0.9615	0.9622	0.9624
Data Set 3	Number of observations	58565	58565	58565
	Number of estimates	6	7	8
	LL(0)	-216342.2	-216342.2	-216342.2
	LL(β)	-9226.5	-8929.7	-8824.4
	Likelihood Ratio Test	414231.5	414825.0	415035.7
	rho-squared	0.9574	0.9587	0.9592
	adjusted rho-squared	0.9573	0.9587	0.9592

4.4 Summary

This chapter proposes a frequency-based path-size formulation as an additional factor to the base path-size value for modeling path overlapping using PSL model in public transport network. The classic path-size formulation and the new frequency-based path-size formulation have been derived with explicit definition of assumptions and properties in the context of public transport network. The main contributions of this chapter is the proposal of new path-size formulation with special treatment to address the unique nature of public transport network. This is the first attempt to capture the effect of overlapped boarding stations on passenger route choice behavior. The formulation were analyzed by model estimation and prediction using smart card data against MNL model and base PSL model with direct application of path-size formulation.

Results indicate that the proposed path-size formulation achieves better estimation and prediction accuracy.

Chapter 5 - Modeling Route Choice Behaviours in Public Transport Network with Multimodal Access and Egress

Travelling in multimodal public transport involves transfers between different public transport modes and services, as well as multimodal access and egress to/from the public transport network. This study expands existing literature by considering the multimodality of public transport trips, not only among different public transport services, but also among different access/egress modes (e.g. walk, bicycle, taxi, car as driver, car as passenger, and motorcycle). While walking and taxi can be assumed to be available to all travellers, this might not be the case with other access/egress modes, particularly car as passenger. A latent choice availability framework is proposed to address the availability issue of access/egress modes. Data from household travel surveys is used to unveil passengers' route choice preferences in the public transport network.

5.1 Introduction

Among the limited literature exploring the preferences of passengers in public transport networks, several studies focus only on the development of multimodal trip assignment without utilizing external preference data for route choice model estimation (Benjamins et al., 2001; Brands et al., 2014; Nuzzolo et al., 2012; Sumalee et al., 2011). The major drawback of this approach is that the model does not specifically reflect the actual route choice behaviour for passengers in the public transport networks.

Other studies use Stated Preference (SP) data (Bradley and Gunn, 1990; Chowdhury et al., 2015; Hess et al., 2008; Vrtic and Axhausen, 2003). The disadvantage of using SP data is the potential bias as the respondents do not experience the actual trip (Louviere et al., 2000).

Few studies use Revealed Preference (RP) data which could provide greater insight into passengers' actual choices in public transport networks. However, when RP data is used, limitations arise to the extension of the network, the modes considered or the alternative paths generated in complex networks. Some studies examine the multimodal route choice behaviour on limited inter-city rail networks (Debrezion et al., 2009; Tsukai and Okumura, 2003; Uges et al., 2002). A few authors study the public transport route choice behaviour between stops/stations in public transport network using smart card data and travel surveys without considering access or egress (Raveau et al., 2011; Schmöcker et al., 2013; Tan et al., 2015).

Literature on multimodal route choice model of public transport passengers in large-scale urban public transport network using RP data has only become available recently (Anderson et al., 2014; Eluru et al., 2012). Eluru et al. (2012) applied a Mixed Logit model in the public transport network in Montreal, Canada while Anderson et al. (2014) estimated a mixed Path Size Logit model in the greater Copenhagen area. However, none of these works explicitly specified the access and egress modes in their public transport trips, and the choice set considered in their works is either a limited number of alternatives from geographic information platforms (such as Google Maps) without checking coverage or directly generated by choice set generation method

without checking the possible biases in the choice set. The effort to reach 100% coverage for all reasonable paths by investigating the fail-to-generated paths helps to correct data error and reduce potential bias in the model estimation as if the choice set generation fails to generate certain type of observed paths, the subsequent route choice model estimation from the generated choice set will not be able to capture the selection preference to such paths.

This chapter goes beyond previous literature in two aspects. Firstly, it considers the multimodality of public transport trips not only among different public transport services, but also among different access/egress modes. These modes can include walking, bicycle, taxi, car as driver, car as passenger and motorcycle. Considering multiple access/egress modes into public transport trips not only enlarges the scale of path choice set, but also imposes additional challenges on modelling due to the complicated availability issue of access/egress modes. Unlike walking and taxi, other access/egress modes are not always available to every passenger, while the availability of access/egress modes directly affects the choice set of paths which passengers are facing when making route choice decisions.

To address the availability issue of access/egress modes, a latent class choice model framework is proposed to capture the influential factors on access/egress mode availability in a latent class model and subsequently model the multimodal public transport route choices using route choice model conditionally on the choice set specified by each latent class. A set of deterministic criteria are firstly used at the choice set generation procedure to exclude paths with behaviourally unreasonable access/egress modes. Then the

latent class model specifically addresses the choice set heterogeneity due to availability of access/egress modes at finer resolution. Consequently, the choice set varies in each latent class by excluding the paths with non-feasible access/egress modes. RP data from household travel surveys is used for model estimation to unveil passengers' route choice preferences in the public transport network in Singapore.

The second aspect in which this study goes beyond previous literature is in terms of choice set generation methods. Efforts have been made to generate choice sets on complete path from origin building to destination building with multi-modal access and egress, with 100% coverage on all reasonable paths in a large and complex multimodal public transport network. As with the multimodality of the paths, requiring 100% coverage results on extremely large choice sets. Additionally, to examine the heterogeneous preferences across passengers in their route choice behaviour, passengers' socio-economic characteristics are taken into consideration in the model framework as well.

This chapter is structured as follows. The multimodal public transport trip is firstly defined in Section 5.2, together with the description of the trip data and network data. Section 5.3 presents the adopted modelling approach for route choice analysis including choice set generation procedure and the latent class route choice model. The estimation results are presented and discussed in section 5.4. In the end, Section 5.6 concludes this chapter.

5.2 Data Set

A complete public transport trip is a multimodal trip, beginning with an access leg from the origin to a stop or station in public transport network, followed by one or several transit legs on public transport system, and ended with an egress leg from public transport system to the destination. The access/egress legs could be travelled not only on foot, but also by bicycle or car, while the transit legs could be traversed among different public transport services such as bus and metro.

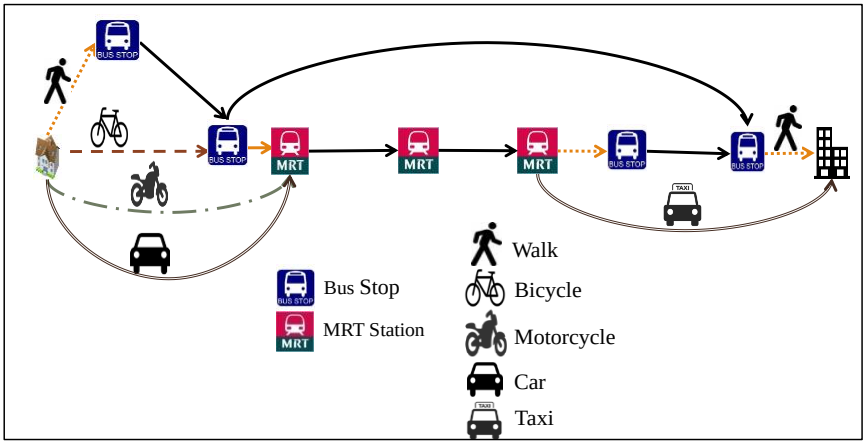


Figure 5-1 An example of multimodal public transport trip in Singapore

In Singapore, walk is the major mode to and from the public transport system, however, other access and egress modes are also observed, including: bicycle, car, taxi, and motorcycle. The public transport system in Singapore contains bus and rapid transit including MRT and LRT. In this study, all these available modes in Singapore are considered. Figure 5-1 shows an example of multimodal trip in Singapore from home to work location.

5.2.1 Travel Data

The travel data used in this study comes from the Household Interview Travel Survey (HITS) conducted by The Land Transport Authority of Singapore (LTA) from June 2012 to May 2013. About one percent of all households in Singapore were surveyed, and all household members above the age of four were asked for their travel details including trip frequency, trip origin, trip destination, trip departure time, travel mode, travel bus line, travel MRT stations, and others(Cheong and Toh, 2010). The trip origin and destination is specified in postcode level where each postcode commonly stands for one building in Singapore. The HITS data includes around 20,307 observations of multimodal public transport trips across 11,575 different surveyed residents, in which the public transportation modes (i.e. Buses, MRTs, LRTs) are the main travel mode. From HITS data, walk is observed to be the predominant access/egress mode (98.71%) for the multimodal public transport trips. Bicycle (0.33%), taxi (0.23%), car as driver (0.10%), car as passenger (0.60%) and motorcycle (0.02%) are also observed as access/egress alternatives for the trips.

A dataset was sampled from all the multimodal trips in HITS for model estimation based on stratified choice-based sampling. As the multimodal trip with non-walk access and egress are extreme minority for trips in the dense public transport network in Singapore, random sampling of all observations would result few observations or even no observation of certain trips with non-walk access or egress. To get a more balanced dataset for model estimation, stratified choice-based sampling is used to retain all the observations on trips with non-walk access/egress, while random sampling is only performed over all

trips with walk access and walk egress. Particularly in this study, all path alternatives in each objective choice set were partitioned into two collectively exhaustive stratums, namely: (i) trips with walk access and walk egress, and (ii) trips with at least one non-walk access/egress leg. Define the population share of each stratum g as W_g, and the sampling fraction of stratum g as H_g. Table 5-1 shows the sampling statistics for the stratified choice-based samples for subsequent model estimation. Note that, the population fraction of each stratum in the complete HITS data is obtained by computing the population fraction of that stratum with respect to the sampling strategy of HITS data from the population.

Table 5-1 Fractions and adjustments in stratified choice-based samples

Stratum	Sample fraction (Hg)	Population Fraction (Wg)	ln(Hg/Wg)
Walk access and Walk egress	80%	98.7%	-0.0912
Trips with at least non-walk access/egress	20%	1.3%	1.1868

This dataset contains 1265 multimodal trips with 1012 trips with walk access and walk egress and another 253 trips with non-walk access or non-walk egress. The origin and destinations of these trips in the data set are distributed over the whole Singapore as shown in Figure 5-2.

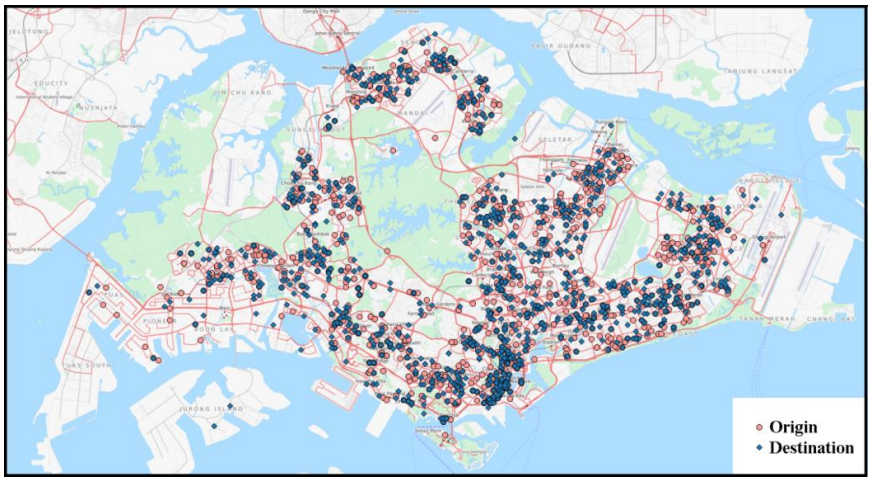


Figure 5-2 Origin and destination distribution for the multimodal trips in the dataset

5.2.2 Network Data

Google Transit data with stop, service line, travel time information for bus, MRT and LRT, service line frequencies, and timetables was used to build up the public transport network and compute link attributes such as in-vehicle travel time, walking time and waiting time. Detailed construction of the public transport network including walk transfer links can be found in Section 3.2. This study extends the public transport network created in Section 3.2 by adding multi-mode access and egress links, with each access link connecting an origin building to a stop/station on public transport network by a specific mode, while each egress link linking up a stop/station to a destination building by a specific mode. For example, a building could be linked by one walk access link and another bicycle link to a nearby MRT/LRT station.

In the interesting work conducted by Uges et al. (2002) on modeling regional train route choice, distance constraints were imposed on access/egress legs such

that the access/egress distance lies between the 10th and 90th percentile of for train trips observed in Dutch National travel survey. In another work considering multiple access modes (walk, car, bicycle, and public transport services) to rail stations, they choose to simplify the network by allowing any three possible rail stations per origin with all four access modes available at one mode per link (Debrezion et al., 2009). In a recent study by Brands et al. (2014) proposed three types of criteria to generate access/egress links to public transport network including maximum distance radius, type of reachable system/station for each zone, and 3) minimal number of stations. They considered walk, bicycle and car as access/egress modes. However, the exact configurations of these criteria related to the examined network were not specified nor justified using empirical data.

The access/egress generation criteria for this work were developed based on the work of Brands et al. (2014) in reference to the real network in Singapore and the empirical observations in HITS data such that the network covers all possible access and egress links. Distance radius by access/egress type, and minimal number of stops/stations by access/egress type were used as criteria to create access and egress links between buildings and stops/stations based on empirical observations and network facilities. The distance radius criterion specifies the maximum access/egress distance radius by each mode between buildings and stops/stations. The specification of all these criteria, shown in Table 5-2, are based on the empirical observations in HITS data and are adjusted according to the investigations of fail-to-generate paths in choice set generation stage. Note that there is no bicycle link from buildings to/from bus stops created

in this study, as in Singapore, bicycles are commonly not able to park at bus stops.

Table 5-2 Generation criteria for access and egress legs from buildings to stops/stations

	Access				Egress			
	to/from MRT/LRT		to/from Bus		to/from MRT/LRT		to/from Bus	
Mode	Radius (km)	Min number of Stations	Radius (km)	Min number of stops	Radius (km)	Min number of Stations	Radius (km)	Min number of stops
Walk	1.2	3	0.75	5	1.2	3	0.75	5
Bicycle	3.0	3	0	0	3.0	3	0	0
Taxi	4.0	5	2	10	7.0	5	3.0	10
Car driver	5.0	7	2	10	5.0	7	2.0	10
Car passenger	7.0	7	4	10	8.0	7	4.0	10
Motorcycle	4.0	5	4	10	4.0	5	4.0	10

5.3 Methodology

In developing a model framework that captures the influence of access/egress mode availability on passenger route choice decisions, this work adopts both deterministic and probabilistic criteria to construct a behaviourally reasonable choice set that passengers are facing when making route choice decisions. A set of deterministic criteria are firstly used at the choice set generation procedure to include paths with all behaviourally reasonable access/egress modes. Then a probabilistic criteria on modal availability is imposed through a latent class model, in which each latent class corresponds to a different availability of

access/egress modes depending on the passenger social-economic characteristics and trip characteristics. The selection of an optimal single path from the choice set is therefore modelled using a multinomial Logit conditionally on the choice set with specified modal availability in the latent class.

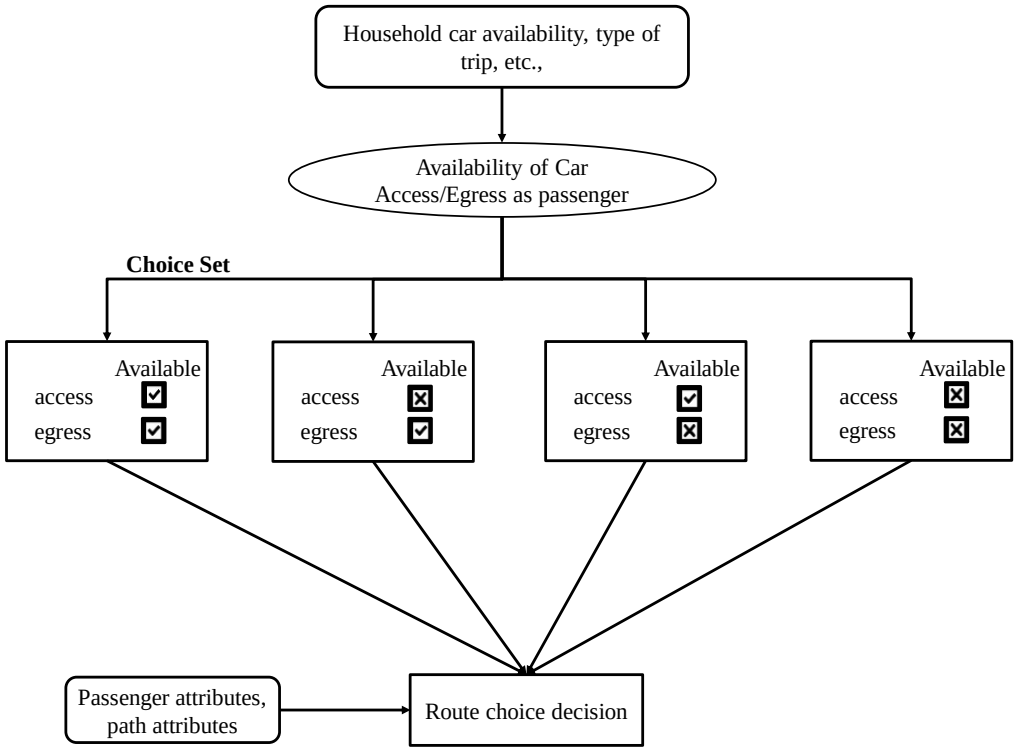


Figure 5-3 Modelling framework of the complete route choice model

The latent class choice model framework proposed in this work is depicted in Figure 5-3. As it is computationally difficult to support fully expansion of availability all access/egress modes, particularly as the number of path alternatives is large, this work only illustrates the most complicated availability of car as passenger using the latent class choice model. However, the model framework developed can be readily extended along other access/egress modes.

Besides, this framework does not require use of latent variables from external sources, although it could be easily extended to incorporate such information.

When availability of car access/egress as passengers is considered in the latent class model, four latent classes should be considered as shown in Figure 5-3. These are: (i) class with both car access and egress as passenger available, (ii) class without car access but with car egress as passenger available, (iii) class with car access but without car egress as passenger available, and (iv) class without car access or egress as passenger available. However, the first class (with both car access and egress as passenger available) cannot be modelled in this application to Singapore, as no such case is observed in HITS data (and therefore the latent class model would not assign travellers to that class). Therefore, in practical application, only the latter three classes are considered in this work.

5.3.1 Choice Set Generation Procedure

Generating the set of alternative paths is very critical to the model estimation results, and it is especially a challenging problem in route choice model (Prato, 2009). In the route choice problem, there is a large number of possible path alternatives for a traveller. This is especially the case in this study as not only different public transport services but also multiple access/egress modes are considered.

In this study, an observation is declared to be matched in the generated choice set if there is a generated path having (i) the same transport modes in the same order for all access leg, egress leg, and transit legs and transfer legs, (ii) exact bus line number for transit legs on bus, and (iii) exact boarding station and

alighting station for transit legs on MRT/LRT. As only bus line number instead of boarding stop and alighting stop is recorded in HITS data for travelling on buses, the coverage is defined at bus line level instead of stop level for bus transit leg on bus.

The following procedure is used to generate the final choice sets of trip alternatives for the origin-destination (OD) pairs of each passenger for model estimation.

- **Step 1 -Mode Availability:** For each OD pair, with respect to the geometry locations of OD and passenger socioeconomic characteristics, check for the access and egress mode availability with deterministic constraints.
- **Step 2 - Boarding Stop and Alighting Stop Feasibility:** For each possible access/egress mode combination, search for all feasible boarding stop and alighting stop pairs on public transport network.
- **Step 3–Choice Set Generation Method:** Given each boarding stop and alighting stop, search for possible paths between boarding stop and alighting stop, and concatenate these paths with their access leg and egress leg respectively to form complete paths from origin to destination.
- **Step 4 – Path Feasibility Check:** Impose path feasibility check for each generated paths to discard any infeasible paths from the choice set.
- **Step 5 - Investigate on Fail-to-generate Paths:** Check whether there is any observed path in the dataset that is not generated. If there is any fail-to-generate path, investigate the failed reason, correct any data error, adjust availability and feasibility configurations in Steps 1 and 2, and fine-tune choice set generation parameters in Step 3. Repeat Step 1 to Step 5 until 100% coverage is reached.
- **Step 6 – Final Result:** Output the final choice set with 100% coverage for subsequent model estimation.

5.3.1.1 Mode Availability

In this study, walk and taxi are assumed to be available to all passengers as both access and egress. When the household possess a bicycle, it is available to a passenger as access mode if the trip begins at home, and as egress mode if the trip ends at home. As bicycle rental is not available in Singapore as a transportation services on public roads, it is reasonable to impose such constraint on bicycle availability as access and egress. For access/egress by car as driver, the household must have a car or van, and the traveller himself should possess a driving licence. As parking is needed for cars, by assuming cars always have to be parked at home at night, this mode is limited to access only if the trip begins at home, and to egress only if the trip ends at home. Similar rules are assumed to apply to motorcycle access/egress.

For car access/egress as passenger, the deterministic constraint is only that the direct distance between origin and destination should be more than 5 kilometres. This constraint is imposed as there is no such observations from the survey data for public transport trips. When the distance between origin and destination is within 5 kilometres and car is utilized as transport mode, car mode is used for the whole trip. The mode availability for car access/egress as passenger will be further addressed at finer details in the latent class choice model.

5.3.1.2 Boarding Stop and Alighting Stop Feasibility

Given the OD geometry location and access/egress mode combination, all boarding stops that could be reached from origin by the specified access mode and all alighting stops that could reach destination by the given egress mode are taken into consideration. All possible boarding stop and alighting stop pairs to

the given OD are enumerated. Two additional constraints are imposed to filter out any infeasible stop pairs: (i) the boarding stop should be different from the alighting stop, and (ii) the total distance of access leg and egress leg should be smaller than the distance between origin and destination.

5.3.1.3 Choice Set Generation Method

Given the boarding stop and alighting stop pair, the Nested Labelling and Link Elimination approach is used to search for possible paths between boarding stop and alighting stop on the public transport network. The Nested Labelling and Link Elimination approach combines the searching heuristics of the Labelling approach as defined in (Ben-Akiva et al., 1984) and Link Elimination approach as used in (Ramming, 2001). It performs the link elimination on the least cost path generated by each label in the Labelling approach. There are four labels used in the Labelling approach where the cost is defined to linear combination of in-vehicle travel time on bus, in-vehicle travel time on MRT/LRT, waiting time, walking time, and transfer penalty, and monetary cost based on the estimation result for a route choice model from boarding stop to alighting stop in Singapore in (Tan et al., 2015).

The parameters on in-vehicle travel time on bus, in-vehicle travel time on MRT/LRT, waiting time, and walking time are adjusted such that Label 1 prefers bus than MRT/LRT and walking than waiting; Label 2 prefers bus than MRT/LRT and waiting than walking; Label 3 prefers MRT/LRT than bus and walking than waiting; and Label 4 prefers MRT/LRT than bus and waiting than walking.

All the stop-to-stop paths generated by this choice set generation method will be concatenated with their respective access leg and egress leg to form complete multimodal public transport paths from origin to destination.

5.3.1.4 Path Feasibility

A path feasibility check is imposed on all generated paths to ensure that all the path alternatives in the choice set for model estimation is feasible and reasonable paths. A feasible path should always start with an access leg from origin to a stop/station on public transport system and ends an egress leg from a stop/station to destination. There should not be any consecutive walk legs along a path and there should be at least one transit leg on bus or MRT/LRT. The maximum number of transit legs on bus, MRT and LRT is constrained to be five. These constraints are empirically consistent to all the observations in HITS data used in this study, as well as all the observations from smartcard data as in (Tan et al., 2015).

5.3.1.5 Investigations on Fail-to-generate Paths

The observation coverage of the initial choice sets for the dataset is 83%. Detailed investigations have been carried out to identify possible reasons for failure of generating observed paths. Five major categories of failed cases were identified and they are:

- **Fail Reason 0:** Reasonable path. These paths are reasonable but failed to be generated due to either the observed access/egress mode is not available or the boarding stop is not included or the Nested Labelling and Link Elimination method failed to generate the paths.

- **Fail Reason 1:**Network input error, including postcode error, bus line stops error, bus line travel time error, travel time error and bus stop latitude-longitude error

- **Fail Reason 2:** Amendable erroneous observations, including paths with reversed access and Egress records, paths with duplicated transfers on MRTs, and paths with additional loop after reaching destination.

- **Fail Reason 3:** Non-amendable erroneous observations: including paths with extremely long walk (>5km) within a short time, paths with non-existing bus line, paths with bus line not running in the reported day of week.

- **Fail Reason 4:** Extremely long detoured trip. These paths are with total travel time 3 times more than the least travel time path in the choice set.

Efforts are made to generate the reasonable paths in Fail Reason 0 by adjusting the mode availability assumptions and stop pair feasibility assumption, as well as fine-tuning cost function for each label in the choice set generation method. Network input errors (Fail Reason 1) and error observations (Fail Reason 2) are corrected. Observations in Fail Reasons 3 and 4 are discarded from the dataset and new observations are re-sampled from the HITS data.

5.3.1.6 Final Result

The final choice sets used for model estimation achieve 100% coverage of observations. Due to the multimodality in both access/egress legs and the transit legs, each OD is having a large number of path alternatives in the choice set. The minimum number of path alternative in the choice set for an OD pair is 30, and the maximum is 11,290. OD pairs with short distances are generally having fewer path alternatives in the choice set while OD pairs with long distances are having more path alternatives in the choice set. Comparing to the size of choice

set for stop-to-stop pairs in Chapter 3 , the size of choice set per OD pair has increased significantly due to the inclusion of multi-mode access and egress legs.

5.3.2 Latent Class Route Choice Model

During the choice set generation process, paths with access/egress by car as passenger are generated for all the OD pairs as long as the direct distance between origin and destination is more than 5 Km. However, when the passenger is about to make a route choice decision, paths alternatives by car as passenger might not be available at that situation. The availability of car access/egress as passenger has been affected a lot by passengers' socio-economic characteristics and trip type. For example, on a home-to-work trip in the morning, passengers without a car in the household will be less likely offered a car access as passenger than passengers with cars in the household. But it is not a strict cut because it is still possible that such a ride could be offered by his/her colleagues or friends. Moreover, it might be relatively easier to take an access ride with family members in the morning than a ride back home with family members in the evening.

In the proposed latent class choice model framework, each latent class corresponds a different availability of car access/egress as passenger. Specially, three latent classes are considered in this work, and they are: (i) class without car access but with car egress as passenger available, (ii) class with car access but without car egress as passenger available, and (iii) class without car access or egress as passenger available. In all three classes, availability of other access/egress modes are assumed to follow the deterministic constraints defined in Section 5.3.1.1. Ideally, another class with both car access and egress as

passenger, referred as full choice set, should be considered. But it is not computationally practical as no observations support such latent class as there is no observation of path with both car access and egress as passenger. Therefore, the conditional probability of an observed path given the full choice set is always smaller than the conditional probability of selecting an observed path given constraint path choice set. Therefore, in this work, such latent class is ignored and only three latent classes are considered.

5.3.2.1 Model Structure

The latent class model is specified as a Logit and the utility of latent class s_n for passenger n may be expressed as follows:

$$U_{S_n} = \boldsymbol{\gamma_s^T Y_{nt}} + \eta_{S_n} \qquad \textit{Eq. (5-1)}$$

where $\boldsymbol{\gamma_s}$ is a vector of coefficients for latent class s; $\boldsymbol{Y_{nt}}$ is a vector of parameters associated with passenger n's characteristics and the trip $\boldsymbol{t}$ characteristics; and η_{S_n} is assumed to be an i.i.d. gumbel distributed random variable across passengers and classes with zero mean and variance $\pi^2/6$. With the utility maximization assumption in Logit, the probability of passenger n facing choice set with latent class s is then denoted as:

$$P(s_n) = \frac{exp(\boldsymbol{\gamma_s^T Y_{nt}})}{\sum_{s' \in \boldsymbol{S}} exp(\boldsymbol{\gamma_{s'}^T Y_{nt}})} \qquad \textit{Eq. (5-2)}$$

The utility of a route alternative is defined as:

$$U_{in} = \boldsymbol{X_{in}^T(\beta + \alpha Z_n)} \qquad \textit{Eq. (5-3)}$$

where $\boldsymbol{X_{in}}$ is a vector of path attribute for path i faced by passenger n, and $\boldsymbol{\beta}$ is a vector of coefficients on path attributes to be estimated. $\boldsymbol{\alpha}$ is a vector of coefficients on passenger socio-economic characteristics $\boldsymbol{Z_n}$. It is a PSL model with path-size included in $\boldsymbol{X_{in}}$. To make this model more generic, the commonly adopted path size value as specified in Eq.(2-3) is used.

Given passenger n facing choice set C_{s_n} with latent class s, the conditional route choice probability is then defined as:

$$P\left(i|C_{s_n}\right) = \frac{exp\left(\boldsymbol{X_{in}^T}(\boldsymbol{\beta} + \boldsymbol{\alpha Z_n})\right)}{\sum_{j \in C_{s_n}} exp\left(\boldsymbol{X_{jn}^T}(\boldsymbol{\beta} + \boldsymbol{\alpha Z_n})\right)} \qquad \textit{Eq. (5-4)}$$

The probability of observing passenger n selecting path i is then:

$$P(i) = \sum_{s_n \in \boldsymbol{S}} P\left(i|C_{s_n}\right) P(s_n) \qquad \textit{Eq. (5-5)}$$

The log-likelihood function to optimize is then:

$$ll(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_n ln\left(\sum_{s_n \in \boldsymbol{S}} P\left(i|C_{s_n}\right) P(s_n)\right) \qquad \textit{Eq. (5-6)}$$

Note that as the coverage of observation is defined at bus line level instead of stop level for bus transit leg on bus, it is possible that an observation could be matched to multiple path alternatives with different boarding/alighting stop on the same bus lines. In this case, the probabilities of all matched path alternative to this observation is summed up to compute the likelihood of this observation for model estimation.

5.3.2.2 Model Estimation

The dataset used in this work is choice-based sampled. Conventional exogenous maximum likelihood (ESML) estimation leads to biased estimation result over choice-based samples. A detailed review on consistent estimation over different sampling method is provided by Ben-Akiva and Lerman (1985). However, when the choice-based sampling is performed over each alternative in the choice set and the choice model is a Logit with full set of alternative specific constants, conditional maximum likelihood (CML) estimation is a consistent estimator (Bierlaire et al., 2008; Manski and Lerman, 1977). In this case, the CML estimation could be performed by adjusting the utility function for choice-based samples and performing ESML over the adjusted utility. Based on the prove, it could be easily extended to show that for stratified choice-based sampling, if the choice model is Logit with full set of stratum specified constants for each choice-based stratum, CML estimation is also consistent. In this work, the route choice model is specified in a Logit model with full set of stratum-specified constants and the CML estimation is performed by subtracting ln(Hg/Wg) in Table 5-1 from the utility of a path.

5.4 Model Specification

Numerous models were estimated with varied utility specifications. Here the key results of three best models with different model specifications are presented in Table 5-3. The first model is the base PSL model, denote as Base PSL. In the Base PSL, the path attributes used to construct the utility function include ten time components, total travel cost, three transfer components, five mode specific constants, and the path size factor to address the correlation due

to path overlapping. Note that for normalization purpose, the mode specific constant for path with both walk access and walk egress is fixed to have zero coefficients and therefore not entering the utility function. Three types of socio-economic attributes are assumed to interact with three different path attributes to address the heterogeneous preference due to different socio-economic background. Binary variable indicating whether the trip is a mandatory trip for education or work is interacted with the total waiting time. Passenger income in 1000 Singapore dollar (S$) is believed to have an impact on passenger's preference over total path cost. Another socio-economic parameter considered in this work is the number of years older than age 55. Here if a passenger is aged 65, he/she is 10 years older than age 55, while if a passenger is younger than age 55, this parameter is set as 0. The selection of age 55 is because while doing a full interactive specifications of all ages with the binary variable for path with at least one leg on MRT/LRT, the data only suggests an influence of age linearly on preference of paths with MRT/LRT for senior passengers over 55 years of old. In the second model – the Improved PSL model, additional parameters are used to interact with the binary variables indicating a path has an egress/egress leg on car as passenger. These interactions are used to introduce penalty to paths with access/egress leg on car as passenger depending on the availability of car in the household and whether the trip is a home-based trip. The third model is the proposed latent class choice model (LCCM). Given a latent class with access/egress availability, the model specification is the same as in the Base PSL model. The latent class membership model is specified to be affected by the availability of car in the household and whether the trip is a home-based trip.

Subsequently, the route choice model is performed conditionally on the heterogeneous choice set which is affected by the availability of access/egress modes specified in each latent class.

Table 5-3 Estimation results for different model specifications

	Base PSL		Improved PSL		LCCM	
	Coefficient	t-stats	Coefficient	t-stats	Coefficient	t-stats
Total waiting time (min)	-0.175	-10.468	-0.177	-10.364	-0.183	-10.440
Binary variable for mandatory trip	-0.075	-3.335	-0.066	-2.853	-0.061	-2.554
Total cost (S$)	-1.593	-9.903	-1.784	-10.208	-1.965	-9.854
Income in 1000S$	0.046	2.098	0.049	2.213	0.058	2.537
Binary variable for path with at least one leg on MRT/LRT	1.964	10.442	1.925	10.136	1.881	9.685
Number of years older than age 55	-0.078	-3.124	-0.076	-2.946	-0.077	-2.801
Access walking time (min)	-0.256	-27.340	-0.254	-26.725	-0.264	-27.021
Egress walking time (min)	-0.230	-26.865	-0.236	-26.886	-0.243	-27.164
Access bicycle time (min)	-0.440	-7.611	-0.431	-7.585	-0.456	-7.804
Egress bicycle time (min)	-0.868	-20.514	-0.882	-20.406	-0.903	-20.861
Access motorized time on taxi/car/motorcycle (min)	-0.375	-10.981	-0.436	-11.656	-0.438	-11.133
Egress motorized time on taxi/car/motorcycle (min)	-0.488	-12.136	-0.312	-7.185	-0.341	-7.887
Total in-vehicle time on MRT/LRT (min)	-0.097	-11.023	-0.102	-11.003	-0.119	-11.975
Total in-vehicle time on bus (min)	-0.167	-21.408	-0.172	-21.379	-0.188	-21.936
Total walk time during transfer (min)	-0.163	-6.730	-0.160	-6.774	-0.166	-6.938
Total number of transfer between buses	-2.958	-20.893	-3.033	-21.014	-3.103	-20.937
Total number of transfer between bus and MRT/LRT	-3.346	-23.782	-3.363	-23.906	-3.447	-23.876
Total number of transfer between MRT/LRT	-0.981	-8.997	-0.992	-8.996	-1.044	-9.227
Binary variable for path has an access/egress leg on bicycle	-3.256	-7.239	-3.234	-7.266	-3.271	-7.233
Binary variable for path has an access/egress leg on taxi	-2.870	-5.425	-2.455	-4.312	-2.045	-3.224
Binary variable for path has an access/egress leg on car as driver	-6.278	-20.069	-6.246	-19.794	-6.278	-19.422
Binary variable for path has an access/egress leg on motorcycle	-7.629	-12.382	-7.696	-12.443	-7.816	-12.435
Path-size factor to address path-overlapping	1.000	n.a	1.000	n.a	1.000	n.a
Binary variable for path has an access/egress leg on car as passenger	-7.354	-33.916	-7.694	-26.454	-5.415	-14.474
Binary variable for path has an access leg on car as passenger			2.437	8.935		
Binary variable for household has vehicle and the trip is from home			-11.296	-19.700		
Binary variable for path has an egress leg on car as passenger			3.847	6.848		
Binary variable for household has vehicle and the trip is to home						
Latent Class of Choice Set:						
Class Constant - car access as passenger is available					-1.409	-3.903
Binary variable for household has vehicle and the trip is from home					3.753	2.491
Class Constant – car egress as passenger is available					-2.298	-5.243
Binary variable for household has vehicle and the trip is to home					1.851	4.325

5.5 Perception and Valuation of Route Choice Behaviours

5.5.1 Estimation Results

As shown in Table 5-3, the estimates of coefficients for all three models have the expected sign, and all the estimates are statistically significant at the 95% significance level.

Comparing the estimates of the three models, parameters related to total travel cost, access/egress motorized time, and binary variables related to car access/egress as passenger have the largest deviation among models. Out of the three models, only the Base PSL model obtains a more negative coefficient on egress motorized time than the coefficient on access motorized time. Consider choice situations where car egress as passenger is not available to passengers but the observed selections are paths with walk access and egress. By putting paths with car egress as passenger into choice set, more paths in the choice set are having positive egress motorized time. However, observations support paths with zero egress motorized time. As more paths with car access as passenger are observed than paths with car egress as passenger, in this case, the estimation process will lead to a more negative estimate on egress motorized time. Therefore, it could be potential estimation bias in the Base PSL model on these two variables when availability of car access/egress passenger is not considered. Similar but shallower effect applies to car access as passenger, since almost three times more selected path with car access as passenger are observed in HITS data than that with car egress as passenger.

Another potential bias could be observed on the estimates on binary variables for path with car access/egress as passenger. In the Base PSL model, binary variable for path has an access/egress leg on car as passenger carries not only the inherent preference on car access/egress as passenger but also largely serves as calibration parameter for mode share of car access/egress as passenger. Coefficient on this binary variable in the Base PSL model is estimated to be -7.354, while in the LCCM model, the coefficient on this binary variable increases to -5.415 and in the Improved PSL model it increases to -5.257 in choice situations where trip starts from home and car is available in the household for passenger. Both the Improved PSL model and LCCM model capture the fact that when a trip starts from home and car is available in the household for passenger, it is more likely for passenger to choose path with car access as passenger.

The positive estimate on the binary variable for household has vehicle and the trip is from home in the Improved PSL model and LCCM demonstrate such effect. Coefficients on total travel cost among three models also vary a lot due to relatively high correlation between access/egress motorized time and total travel cost. This is particularly the case in Singapore as the total cost for most of the trips on public transport network has slight variation at low cost between 70 cents to 2.5 dollars while the adoption of motorized vehicle such as cars or taxi is likely to boost up the total cost significantly.

Table 5-4 summarizes for each model the number of observations in the dataset, the number of parameters estimated, the final log-likelihood, the log-likelihood

ratio test, the rho squared and the adjusted rho squared, the Bayesian information criterion (BIC) (Greene and Hensher, 2003) and the correct prediction. The BIC formulation is adopted the same as in Greene and Hensher (2003) while the correct prediction is computed as the percentage of observed trips that are predicted with the highest probability in the choice set using estimated coefficients. As is apparent from Table 5-4, the LCCM model outperforms the base PSL model and the improved PSL model in terms of all test statistics. Besides, the LCCM model is behaviourally more appealing than other two models. If certain access/egress modes are not available in the choice situation, it is behaviourally more reasonable to exclude these paths from the choice set than to penalize them as in the improved PSL model or to totally ignore such availability issue as in the base PSL model.

Table 5-4 Summary of statistics for different model specifications

	Base PSL	**Improved PSL**	**LCCM**
Number of observations	1265	1265	1265
Number of parameters	23	26	27
Final LL	-3246.10	-3140.01	-3089.17
Likelihood_Ratio	10534.47	10746.65	10848.33
rho_squared	0.619	0.631	0.637
Adjusted rho_squared	0.616	0.628	0.634
BIC	-3264.42	-3157.74	-3106.61
Correct Prediction	0.500	0.504	0.532

5.5.2 Interpretation on Mode Availability

Of particular interest is the result for latent class membership. The estimates for the class membership model are listed in the last four rows in Table 5-3 under the LCCM model. As the class constant for choice set without car access or egress as passenger is normalized to zero, the negative coefficients of class constants for choice set without car access as passenger and choice set without car egress as passenger indicate when there is no car in the household and the trip is not a home-based trip, car access or egress as passenger are generally less likely to be available. When the household has car and the trip is from home, the chance of having car access as passenger increases significantly while when the household has car but the trip is to home, the change of having car egress as passenger increases. Here, to which degree the availability of car access/egress as passenger is not as clear-cut, by just looking at the estimates. The class membership probability under different scenarios are then computed in Table 5-5 to provide clear insights on the availability of car access/egress as passenger.

Table 5-5 Class membership probability for the LCCM under different scenarios

Scenario	Class without car access but with car egress as passenger available	Class with car access but without car egress as passenger available	Class without car access or egress as passenger available
Household has car and the trip is from home	0.87%	90.45%	8.68%
Household has car and the trip is to home	33.93%	12.98%	53.09%
Household has no car and the trip is not home-based	7.47%	18.17%	74.36%

When there is no car available in the household and the trip is neither from home nor to home, the probability that there is no car access or egress as passenger available to that passenger is around three quarters. However, when there is car in the household and the trip is originated from home, the probability that car access as passenger is available increases significantly to more than 90%, while when the trip is destined at home, the probability of car egress as passenger increases to around 50%. Such difference is behaviourally reasonable as it is by and large easier for household members to departure from home together in a household car in the morning than to get back home together in the evening. They might not finish the work at the same time or there are some other errands to do after work in the evening.

5.5.3 Route Choice Behavioural Interpretation

The estimation results of the LCCM also demonstrate the effects of various path attributes and passenger socio-economic characteristics on the route choice decision-making process. In general, passengers are more sensitive to waiting time than walking time and in-vehicle travel time. Especially when passengers are on mandatory trips for work or education, they value total waiting time the most negatively. The in-vehicle travel time in MRT/LRTs is much more positively perceived than that in buses. This is because MRT/LRTs in Singapore run much faster than buses at much higher frequencies. This is also consistent with findings in the literature (Anderson et al., 2014; Eluru et al., 2012; Tsukai and Okumura, 2003; Van der Waard, 1988).

The binary variable for path with at least one leg on MRT/LRT has positive sign indicating passengers prefer to travel on MRT/LRT than pure buses. However, as indicated by the negative estimated coefficient of number of years older than 55 against the binary variable that a path has at least one leg on MRT/LRT, when senior passengers get older and older, they tend to favor MRT/LRT less and less. Table 5-6 lists the estimated coefficients of the LCCM model for senior passengers at different ages against the binary variable for path with at least one leg on MRT/LRT. For passenger younger than age 55, they have the strongest favor on path with MRT/LRT. However, this preference over MRT/LRT decreases with the increase of age among senior passengers. When people gets more than 80 years, they favor path with MRT/LRT less than path purely on buses.

Table 5-6 Preferences on path with at least one leg on MRT/LRT

Passenger Age	**<=55**	**57**	**60**	**65**	**70**	**75**	**80**
Binary variable for path with at least one leg on MRT/LRT	1.881	1.728	1.498	1.115	0.733	0.350	-0.033

Mode specific constants for non-walk access/egress modes are all negative and statistically significantly different from zero. As the mode specific constants for path with walk access and walk egress is normalized to zero, a negative coefficient of other mode specific constant suggests that passengers in general less likely to choose this mode than walking when it is available. The estimation results indicate passengers prefer to reach public transport system by foot than

by other modes. It is behaviourally reasonable in Singapore given its intense public transportation network with good accessibility by walk. Moreover, bicycle is not convenient to ride on public road nor to park at bus stops and MRT stations in Singapore, while taxi always requires additional monetary cost and car/motorcycles are always subject to complicate availability.

5.6 Summary

In summary, this chapter developed a latent class choice model framework to tackle the challenges and complexity in modelling complete public transport route choice decisions with multi-mode access and egress in the large-scale urban public transport network in Singapore. The contributions of the work presented in this chapter are four-fold.

First, to the authors' knowledge, this is the first attempt to analyse complete public transport trips with multimodal access from the origin at building level and multimodal egress from to destination in large-scale urban public transport networks. Six possible access/egress modes are considered including walk, bicycle, taxi, car as driver, car as passenger and motorcycle, while all public transport services in Singapore, namely bus, MRT and LRT, are included.

Second, it proposes a latent class choice model framework to address the availability issues of different access/egress modes in modeling route choice behaviour of complete public transport trips. The latent class model utilize a Logit model structure to examine the availability of access/egress modes depending on the passenger social-economic characteristics and trip characteristics. Conditionally on the choice set with specified modal availability

in the latent class, the route choice decision is modelled using Logit model considering both path attributes and passengers socio-economic characteristics.

Third, efforts were also made to generate choice sets with 100% coverage on all reasonable paths in the large and complex multimodal public transport network in Singapore. It helped to correct data errors and avoid potential bias in the model estimation due to the deficiency of choice set in covering certain type of observed paths. To tackle the computational difficulty due to the extreme dominancy of walking as access/egress to public transportation network in Singapore, a stratified choice-based sampling and conditional maximum likelihood estimation was then adopted.

Finally, the estimation results based on passengers' real route choice selections collected in HITS data in Singapore illustrated the effect of household car ownership and trip type on availability of car access/egress as passengers, and unveiled passengers' heterogeneous route choice preferences with respect to different socio-economic characteristics over different path attributes such as different types of transfers, mode of access/egress, and different travel time components including in-vehicle travel time, waiting time for different public transport service modes as well as access and egress time via multiple travel modes.

Chapter 6 - Evaluating Rapid Transit Network Performance: an Application of Route Choice Model Combining Survey and Smart Card Data

This chapter presents an application using route choice model to assess the network performance of the rapid transit network in Singapore. Despite smart card in Singapore captures almost all route choice selections from initial boarding stop to final alighting stop, it does not record transfer information on rapid transit network. This chapter compliments smart card data with the completeness of HITS data by modeling passengers' route choice behaviour on rapid transit network using HITS data. Based on the complete travel demand recorded on rapid transit network in smart card data, it estimates the passenger flows on the rapid transit network, identifies the transfer demand and predicts the probability of fail-to-board and fail-to-seat at each transfer station. It demonstrates how to apply passenger route choice model and get realistic estimation of passenger flows in rapid transit network in Singapore.

6.1 Introduction

To better assess the performance of rapid transit network in Singapore, realistic estimation of passenger flows on rapid transit network is essential. Despite the smart card system in Singapore captures every boarding and alighting information on rapid transit network both spatially and temporally, they are fragmentary in terms of exact path selection. Passengers are only required to tap in at initial boarding station and tap out at the final alighting station on MRT/LRT network. There is no record of transfer stations in rapid transit

network. This drawback of smart card data prevents practitioners from using smart card data to accurately analyze the network performance of rapid transit in Singapore directly as there is no information on which route passengers choose. This drawback is also the reason we dropped all the observations of pure MRT/LRT trips when estimating a stop-to-stop route choice model using smart card data in Chapter 3 .

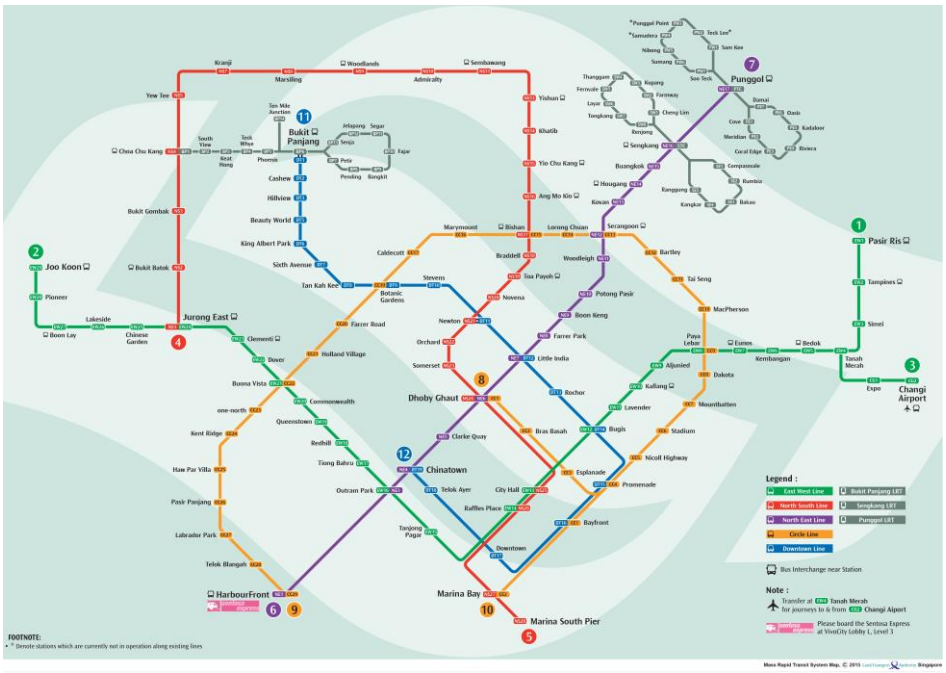


Figure 6-1 The map of rapid transit network in Singapore in early 2013 (source: Land Transport Authority, Singapore)

This chapter presents an application to evaluate the network performance of rapid transit systems by modeling passengers' route choice decisions using HITS data and subsequently estimating passenger flows on the rapid transit network given complete travel demand recorded in smart card data. As is evident from the map of the rapid transit network in Singapore in Figure 6-1, there are more than 16 possible transfer stations with quite a number of station-

to-station pairs that can be travelled by more than one feasible path. The analysis in this chapter is based on more than 2 million trips made in a typical weekday, more than 75% of which must include transfers.

This chapter is structured as follows. The datasets used for route choice estimation and assessment of network performance are firstly introduced in Section 6.2, together with the description of choice set generation method to generate feasible paths set for each station-to-station pair. Section 6.3 presents the estimation of passengers' route choice selections particularly on rapid transit network. In section 0, the assessment of network performance based on the estimated route choice model is presented given complete travel demand recorded in smart card data. In the end, Section 6.5 summarizes this chapter

6.2 Data

6.2.1 Datasets

The data set for route choice model estimation is extracted from HITS 2012 data. It includes all pure MRT/LRT trips and MRT/LRT segments from trips using both rapid transit and buses recorded in HITS. The total number of valid trips are 11,154. To cater for data records in smart card data, all origin and destinations are discarded. Instead, the first boarding MRT/LRT station is treated as origin while the last alighting MRT/LRT station is treated as destination.

The demand dataset is retrieved from smart card data from Aug 2013. Despite that the data is collected few months after HITS 2012 data has been completed, there is no new rapid transit line or station put in operation during these months.

The dataset is retrieved from 05 Aug 2013, which is a Monday. All MRT/LRT trips and trip segments on MRT/LRT is retrieved. The whole dataset contains 12,951 unique station-to-station pairs in the rapid transit network covering 2,219,637 observations.

6.2.2 Choice Set Generation

The branch & bound approach is adapted in this section to search for all possible paths in the rapid transit network of Singapore. It applies to both estimation dataset and demand dataset. Although the branch & bound approach is found not suitable for generating path choice set on the whole public transport network, as discussed in Chapter 3 , it is quite feasible for the rapid transit network as the number of possible transfer station is limited.

Table 6-1 Description of implemented constraints for the branch & bound approach for the rapid transit network in Singapore

Constraints	Type	Constraint Description
1	Logical	Path is loopless
2	Logical	Transfer must be at interchange station
3	Feasible	Maximum allowable number of transfer (<=4)
4	Behavioural	No search of path with three more transfers than least transfer path

Table 6-1 depicts the constraints implemented for branch & bound approach particularly for the rapid transit network. In this work, the least transfer path is firstly identified using shortest path algorithm and it serves as a key reference for the constraint implemented for the branch & bound approach in this work.

The constraint that transfer must be at interchange station is a feasible yet extremely useful constraint for this network. It helps to effectively limit the number of possible search branches during path searching and therefore speeds up the whole choice set generation process. The maximum computational time to search for paths with five transfers is below 10 seconds.

The maximum allowable number of transfer for branch & bound approach is set to be four. This is a valid constraint as in Singapore, passengers could reach any station in the network within four transfers no matter which station they board. All observation in HITS data also supports such constraint empirically as there is no observations of MRT/LRT trips with more than four transfers.

The last constraint is to stop searching path with three more transfers than the least transfer path. This is a reasonable behavioural assumption as number of transfer is always weighted as a significant disutility for route choice models. In our previous work on both stop-to-stop route choice model and complete route choice model, the rate of substitution between number of transfer and total in-vehicle travel time in Singapore is always larger than 10 minutes per transfer. Therefore, it is reasonable to assume passengers will not select paths with three more transfers than least transfer path in rapid transit network. This assumption is also supported by the observations in HITS data empirically.

6.3 Route Choice Model on Rapid Transit Network

The path size logit model specified in Chapter 3 , is adopted for modeling passengers' route choice behaviour in rapid transit network.

Total in-vehicle travel time, total walking time at transfers, and total waiting time, number of stations passing-by are time related components considered in this route choice model. For total walking time, we further differentiate it by total walking time at transfer station within city center and outside of city center. Here if a path transfers at any station inside the core central business district in Singapore, the walking transfer time at this station is counted as walking time within city. As there is no information on dwell time at each station, the total number of stations passing by is used to capture the effect of dwelling time on passenger's route choice decisions.

Two types of transfers are considered in this work, namely, transfer without level change and transfer with level change. If a transfer is conducted at the same platform without going ups and downs, this transfer is considered as a transfer without level change, while if passengers need to go up/down to another platform, it is considered as a transfer with level change.

One alternative specific constant is used in this model, indicating whether the path utilizes the new MRT line - Circle Line (at year 2012) for transfer is also included. It is necessary as when HITS 2012 data was collected, the circle line just put into operation few months ago. This constant is therefore included to capture the behaviour assumptions that passengers are not that familiar with Circle Line comparing to other existing lines. Last but not least, path-size is included to capture the correlation among paths due to path overlapping. Note that the total monetary cost is discarded from this model as fare is fixed for each MRT/LRT station pair regardless which route passengers choose.

The estimation results is tabulated in Table 6-2 below. All the parameters are statistically significant with desired sign.

Table 6-2 Estimation results of route choice model in rapid transit network

Variables	PSL		
	Coefficient	**Std.Err.**	**t-stats**
Total walk time during transfer at city center (min)	-0.769	0.040	19.313
Total walk time during transfer outside of city center (min)	-0.415	0.053	7.892
Total in-vehicle travel time (min)	-0.145	0.014	-10.470
Total waiting time (min)	-0.226	0.030	7.579
Total Number of transfer without level change	-3.519	0.161	-21.852
Total Number of transfer with level change	-4.522	0.179	-25.265
Constant indicate path transferring at station along Circle Line	-0.546	0.186	-2.928
Total number of stations passing by	-0.215	0.029	-7.475
Path-size	0.591	0.068	8.657
Summary of Statistics:			
Number of observations	11154		
LL(0)	-24320.5		
LL(β)	-1792.9		
Likelihood Ratio Test	45055.2		
rho-squared	0.926		
adjusted rho-squared	0.926		

6.4 Assessing Network Performance

In this section, a simulation procedure is firstly developed to estimate the passenger flow in rapid transit network based on the estimated route choice model with total travel demand recorded in smart card data as described in section 6.2.1. The key idea of this Monte-Carlo simulation procedure lies in simulating passengers' 1) path selection, 2) vehicle boarding and 3) vehicle

alighting in an agent-based approach and distributing the load into each vehicle runs, service line, and station.

6.4.1 Simulation Procedure

The simulation procedure is elaborated as follows:

- Step 0: Initialization. Set passenger index $n = 1$, set boarding demand $N_{r,s}^{board} = 0$; alighting demand $N_{r,s}^{alight} = 0$; and passing through demand $N_{r,s}^{pass} = 0$ for $r \in$ all service line run at each traversing stop s.

- Step 1: For passenger $n = 1$ in the demand dataset, retrieve its path choice set and update path attributes with respect to departure time per time interval.

- Step 2: Compute path utility for all paths and path selection probabilities based on estimated PSL model.

- Step 3: For passenger n, simulate his/her path selection following the path selection probability computed from the route choice model. Denote the departure time for passenger n as T_n^{dep} and the selected path k_n in the form as sequence of transfer stations $\{S_1^k, S_2^k, \dots, S_{K+1}^k\}$ and service line to board in sequence $\{l_1^k, l_2^k, \dots, l_K^k\}$ where K is the index for the last transit leg along path k and S_{K+1} is therefore the final alighting station.

- Step 4: Set index $q = 1$, and set current time $T_n = T_n^{dep}$

- Step 5: Update $T_n = T_n + T_{k,q}^{txf}$ where $T_{k,q}^{txf}$ denotes the transfer walk time from previous alighting stop S_{q-1}^k to current boarding stop S_q^k along path k_n. Note that for the first leg, the transfer walk time is always set to be the time from station gantry to boarding platform. Check for

the first arriving train service-run, at stop S_q^k for service line l_q^k, denote it as r_{n,s_q^k}

- Step 6: Check boarding and seating feasibility. Denote the train capacity for this train service-run as $z_{r_{n,s_q^k}}$ for service line run r_{n,s_q^k}, and its seat capacity as $z_{r_{n,s_q^k}}^{seat}$.

- Step 7: Simulate boarding at train service-run r_{n,s_q^k} and alighting at station S_{q+1}^k. Update T_n to the arrival time of train service-run $line_{q,r}$ at station S_{q+1}. Update boarding demand to service-run r_{n,s_q^k} at stop S_q^k : $N_{r,s_q^k}^{board} = N_{r,s_q^k}^{board} + 1$; Update alighting demand to service-run r_{n,s_q^k} at stop S_{q+1}^k : $N_{r,s_{q+1}^k}^{board} = N_{r,s_{q+1}^k}^{board} + 1$; update passing through demand to service-run r_{n,s_q^k} for all stations s between S_q^k and S_{q+1}^k along l_q^k : $N_{r,s}^{pass} = N_{r,s}^{pass} + 1$

- Step 8: Go to Step 9 if $q = k_S$ (the last boarding stop), otherwise, increment q and move back to Step 5

- Step 9: Stop if $n = N$ (total number of passengers), otherwise, increment n and move back to Step 1

The inputs to the simulation procedures are the total travel demand with departure time in demand dataset and the estimated route choice model, while the output of the simulation procedure is the time-dependent traffic load to each service run at each station in rapid transit network. It outputs the flow rate for each service line in hour and for each station, as well as the boarding rate, alighting rate and transfer rate at each station in each hour.

6.4.2 Network Performance Analysis

Figure 6-2 depicts the flow rate to each service line and the boarding rate to each MRT/LRT station at different time of day. Denote travel leg from one station to its subsequent station along the same service line as a link, the width of the link is set in proportional to the passenger flow on that link, while the radius of the circle is in proportional to the total number of passengers boarding at that station. There are four maps in this figure, each at different time of day.

The first map shows the flow rate at 7AM in the morning. High boarding rate at stations outside of central business district (south center) and the flow rate for every MRT line is very high. The second map shows the network performance at 10AM. Comparing to the map for 7AM, both boarding rate and flow rate has decreased significantly. The third map is for 6PM in the evening. Heavy load is observed at central business district, and similar to the first map, high flow rate is observed at every MRT line. The last map indicates the decreased travel demand at 9PM. Tail-effect is observed for afternoon peak as the decrease of travel demand from 6PM to 9PM is much slower than that from 7AM to 10 AM.

In all four maps, the flow rate from station Haw Par Villa (CC25) to station Harbour Front (CC29/NE1) along circle line, is not very high. It is partially because the total demand to these station is not high due to the low residential density and low commercial density. It also indicates this part of MRT line is not much utilized for transferring purpose, as the circle line is not a full circle yet. Currently it does support short travel from CC29/NE1 to stations such as CC1 and CC2 in the center business region.

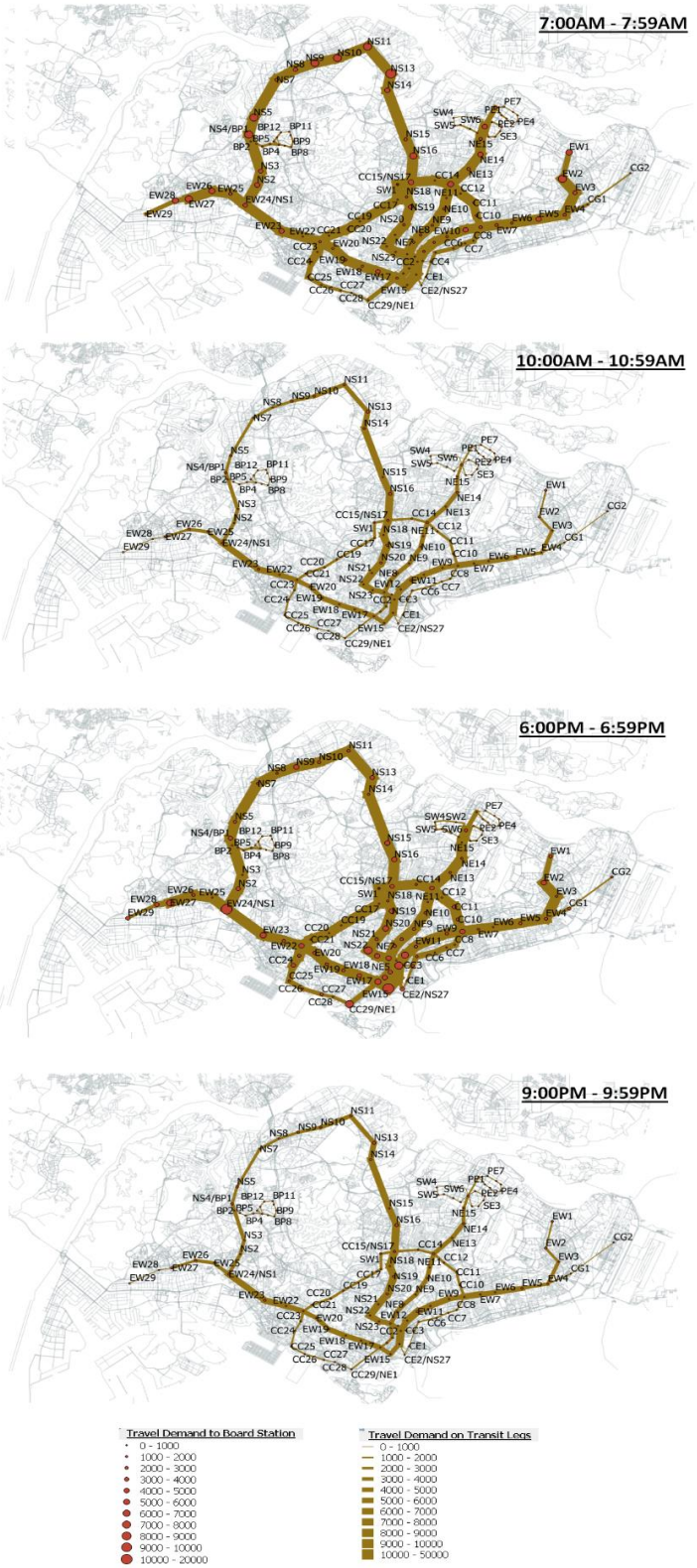


Figure 6-2 Flow rate along transit link and boarding rate to station at different time of day

Another important output from the simulation is the transfer rate at each station per hour. Figure 6-3 presents the bar plots of transfer rate to three stations at different time of day. These three stations have the highest transfer rate among all transfer stations.

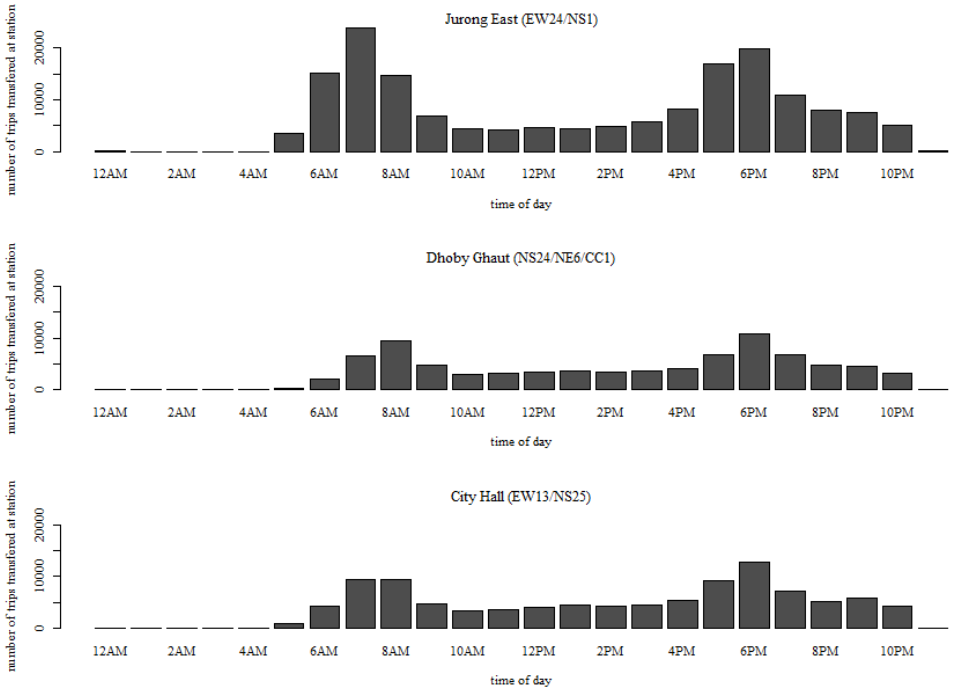


Figure 6-3 Transfer rate to stations at different time of day

The transfer demand to Jurong East (EW24/NS1) is no doubt the highest, much higher than the second and third highest transfer demand at Dhoby Ghuat (NS24/NE6/CC1) and City Hall (EW13/NS25). The transfer demand follows the travel pattern of total travel demand with AM peak in the morning and PM peak in the afternoon. In particular, the transfer demand to all transfer stations are shown in Figure 6-4 where the radius of the circle is proportional to the total transfer demand to that station at 7:00AM to 7:59AM.

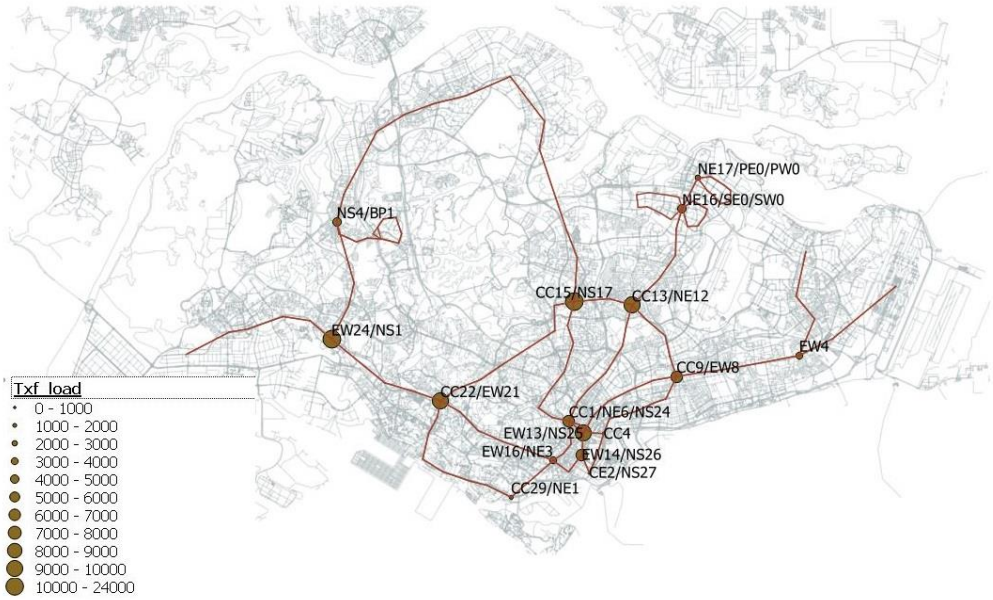


Figure 6-4 Transfer demand to all transfer stations at 7AM

Based on the recent work by Schmöcker et al. (2011), the fail-to-board probability q_s^{FB} and fail-to-seat q_s^{FS} probability has been computed for transfer station s with respect to train capacity z_r for service line run r, train seat capacity z_r^{seat} for service line run r, boarding demand $N_{r,s}^{board}$, alighting demand $N_{r,s}^{alight}$, and passing-through demand $N_{r,s}^{pass}$ as follows:

$$q_s^{FB} = \begin{cases} 0 \\ 1 - max\left(0, min\left(\frac{z_r - N_{r,s}^{pass}}{N_{r,s}^{board}}\right)\right) \end{cases} \qquad \textit{Eq. (6-1)}$$

and

$$q_s^{FS} = \begin{cases} 0 \\ 1 - max\left(0, min\left(\frac{z_r^{seat} - N_{r,s}^{pass}}{N_{r,s}^{board}}\right)\right) \end{cases} \quad \textit{Eq. (6-2)}$$

Figure 6-5 shows the estimated time-dependent fail-to-board probability and fail-to-seat probability for the busiest transfer station Jurong East (EW24/NS1), which is the interchange station between North-South line and East-West line. The train loading capacity is set to be 1500 passengers per train and seat capacity is 270 for both North-South line and East-West line. Note that although the safety limit capacity for trains on North-South line and East-West line is 1920, the maximum loading capacity at full operation is around 1500 (Ministry of Transport, 2010; sgwiki, 2013). It shows, during morning peak, fail-to-board probability on East-West line towards Changi is high, while during afternoon peak, the fail-to-board probability on North-South line towards Woodlands is high. This is due to the heavy travel demand for home-work trips starting from west side to city center in the morning and similarly heavy travel demand for work-home trips from city center to west side in the evening. There are large residential areas around stations along East-West line towards Joo Koon after Juron East, such as Boon Lay and Lakeside, and along North-South line towards Woodlands, such as Bukit Batok and Chao Chu Kang. Due to the high travel demand and limited seat availability, the fail-to-seat probability is very high at Jurong East in both AM peak and PM peak, especially on East-West line towards Changi and on North-South line towards Woodlands.

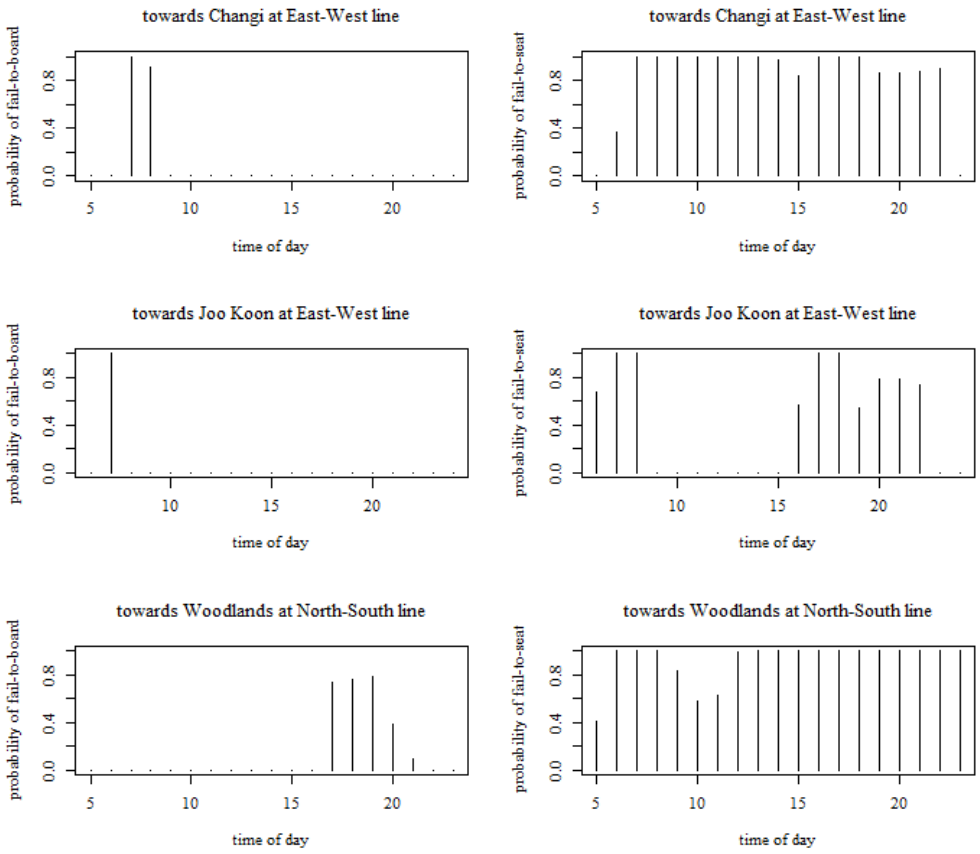


Figure 6-5 Fail-to-board and fail-to-seat probabilities for station Jurong East (EW24/NS1)

6.5 Summary

In summary, this chapter demonstrates an application of route choice model in evaluating the rapid transit network performance in Singapore using both smart card and survey data. Smart card and survey data complements each other in terms of trip coverage and trip completeness respectively. This work firstly adopts survey data to estimate passengers' path selection by modelling realistic passengers' route choice behaviour in rapid transit network. Path attributes such

as number of transfers with and without level change are included with special consideration to rapid transit network. Based on estimated passengers' path selection behaviour, it then utilizes the full travel demand records in smart card data to estimate the passenger flow rate in rapid transit network, identify the transfer demand, and predict the probability of fail-to-board and fail-to-seat at each transfer station.

The contribution of this chapter is its practical importance and translational impact on the real application. The application addressed the data supplementary issue existed in current travel data collected in Singapore as it complements the missing transfer station in smart card data by modeling passengers' route choice behaviour in rapid transit network using survey data in HITS. The simulation procedure could be easily implemented by system operator and practitioners to identify realistic passenger flows at each service line and MRT/LRT station.

Chapter 7 – Conclusions

7.1 Concluding Remarks

It is essential to identify relevant factors that affect passengers' travel behaviour in a realistic manner as passengers' travel behaviour is highly correlated with the performance of public transport systems. With this core objective, this thesis has modeled and analyzed passengers' route choice behaviour in multimodal public transport network in Singapore from different perspectives using different data sources.

Chapter 3 and Chapter 4 presents route choice modeling using smart card data. Chapter 3 investigates passengers' stop-to-stop route choice behaviour on public transport network in Singapore using smart card data. To examine the effectiveness of different choice set generation methods, a comprehensive evaluation framework is proposed including computational performance, three types of coverage test, detailed analysis of fail-to-generate paths, and composition evaluation based on four criteria. A stop-to-stop route choice model is estimated based on final choice set with 100% coverage against one data set and the model is validated by examining the prediction performance against another two datasets using the estimated parameters. Based on the same dataset, Chapter 4 is devoted to formulate a path-size definition in consideration of the special characteristics of public transport network. When passengers make route choice decisions, path alternatives in the choice set are often highly correlated due to various reasons such as path overlapping. This part of the

thesis proposes and examines a new definition of path size in consideration of the nature of public transport using smart card data.

Chapter 5 utilizes travel survey data and is devoted to model passenger route choice behaviour from origin to destination. Travelling in multimodal public transport network involves transfers between different public transport modes and services, as well as multimodal access and egress. While walking and taxi can be assumed to be available to all travellers, this might not be the case with other access/egress modes, particularly for access/egress in car as passenger. A latent choice availability framework is therefore proposed to address this availability issue of access/egress in car as passenger.

Chapter 6 combines smart card data and survey data and presents an application using route choice model to assess the network performance of Singapore's rapid transit network. Despite smart card system in Singapore collects almost all the public transport trips, but it does not contain transfer information on rapid transit network. The comprehensive coverage of smart card data is complemented with the completeness of travel survey data by modeling passengers' route choice behaviour in rapid transit network using HITS data. Based on the complete travel demand recorded in smart card data, it estimates the passenger flows on each train service line, identifies the transfer demand and predicts the probability of fail-to-board and fail-to-seat at each train station.

7.2 Contributions

The original contribution of the thesis is three-fold.

Empirically, it evaluates different choice set generation methods under a comprehensive evaluation framework (Chapter 3), and it identifies relevant

factors that affects passenger route choice behaviour on public transport network using revealed preference data from smart card and surveys (Chapter 3 & Chapter 5). To the authors' knowledge, this is the first attempt to analyse complete public transport trips with multimodal access from the origin at building level and multimodal egress from to destination in large-scale urban public transport networks. Efforts have been made to generate choice sets on complete path from origin building to destination building with multi-modal access and egress, with 100% coverage on all reasonable paths in a large and complex multimodal public transport network. Heterogeneous preferences across passengers in their route choice behaviour has been carefully analysed by taking passengers' socio-economic characteristics and various path attributes into consideration.

Methodologically, it formulates a new path-size definition to address the correlation during path-overlapping on public transport network (Chapter 4), and it proposes a latent class choice model framework to address the availability issues of different access/egress modes in modeling route choice behaviour of complete public transport trips (Chapter 5). The proposed new path-size formulation addresses correlation due to path overlapping by considering the unique characteristics of public transport network. This is the first attempt to capture the effect of overlapped boarding stations on passenger route choice behavior. In Chapter 5, the proposed latent class model framework examines the availability of access/egress modes depending on the passenger social-economic characteristics and trip characteristics. The route choice decision is therefore modelled using Logit model considering both path attributes and passengers socio-economic characteristics, conditionally on the choice set with

specified modal availability in the latent class. The proposed approach has proved to be more accurate and behaviourally appealing in capture passenger's route choice preferences.

Last but not least, the application on the rapid transit network in Singapore (Chapter 6) has translational impact on the current travel data as it complements the missing transfer station in smart card data. It demonstrates how to apply route choice model and get realistic estimation of passenger flows on train stations and service lines. The application has addressed the incompleteness of current travel data collected in Singapore as it complements the missing transfer station in smart card data by modeling passengers' route choice behaviour in rapid transit network using survey data in HITS. The simulation procedure could be easily implemented by system operator and practitioners to identify realistic passenger flows at each service line and MRT/LRT station.

7.3 Future Directions

The following discussion focuses on some promising future research directions related to extensions and improvements of the discrete choice models in the context of passenger pre-trip route choice modeling.

A major source of estimation and predication errors in route choice models is due to passenger failure to consider all feasible alternatives. To identify relevant paths for a consideration set out of complete choice set is helpful in both model estimation and prediction. The work on evaluation of choice set generation methods in this thesis provides empirical evidence on it. Instead of taking all feasible choices into consideration, passengers are observed to have a small consideration route choice set when making route choice decisions. However,

not much research has been done on identifying the consideration set for route choice models, either on road network or public transport network. Current available approaches are extremely computational complex and have never been applied on real network.

Another extension of the current work is to expand route choice modeling from selection of paths to selection of route choice strategies. In this thesis, although the treatment to common line problems introduces selection strategy of service lines to certain extent, the path definition in the implemented route choice models is still path based, which does not fully involve path selection strategy. Hyper-path concept is widely considered for path selection strategy on public transport network in traffic assignment models. It is of interest to collect data particularly on passengers' path selection strategies, for example, hyper-paths, and investigate how passengers evaluate these path strategies. Route choice models based on path strategies can be estimated with comparison to path-based route choice model.

References

Abdelghany, K., Mahmassani, H. (1999) Multi-Objective Shortest Path Algorithm for Large Scale Intermodal Networks. *INFORMS, Philadelphia, Fall*.

Agard, B., Morency, C., Trépanier, M. (2006) Mining public transport user behaviour from smart card data. *Proceedings of 12th IFAC Symposium on Information Control Problems in Manufacturing-INCOM*, pp. 17-19.

Alfred Chu, K.K., Chapleau, R. (2008) Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record: Journal of the Transportation Research Board* 2063, 63-72.

Anderson, M.K. (2013) Behavioural Models for Route Choice of Passengers in Public Transport. Technical University of Denmark.

Anderson, M.K., Nielsen, O.A., Prato, C.G. (2014) Multimodal route choice models of public transport passengers in the Greater Copenhagen Area. *EURO Journal on Transportation and Logistics*, 1-25.

Azevedo, J., Santos Costa, M.E.O., Silvestre Madeira, J.J.E., Vieira Martins, E.Q. (1993) An algorithm for the ranking of shortest paths. *European Journal of Operational Research* 69, 97-106.

Bagchi, M., White, P. (2005) The potential of public transport smart card data. *Transport Policy* 12, 464-474.

Bastin, F., Cirillo, C., Toint, P.L. (2006) Application of an adaptive Monte Carlo algorithm to mixed logit estimation. *Transportation Research Part B: Methodological* 40, 577-593.

(2007) *Quick response freight manual II*.

Bekhor, S., Ben-Akiva, M.E., Ramming, M.S. (2006) Evaluation of choice set generation algorithms for route choice models. *Annals of Operations Research* 144, 235-247.

Bekhor, S., Ben-Akiva, M.E., Scott Ramming, M. (2002) Adaptation of logit kernel to route choice situation. *Transportation Research Record: Journal of the Transportation Research Board* 1805, 78-85.

Ben-Akiva, M., Bergman, M., Daly, A.J., Ramaswamy, R. (1984) Modeling inter-urban route choice behaviour. *Proceedings of Proceedings of the 9th International Symposium on Transportation and Traffic Theory, VNU Press, Utrecht*, pp. 299-330.

Ben-Akiva, M., Bierlaire, M. (1999) Discrete choice methods and their applications to short term travel decisions. *Handbook of transportation science*. Springer, pp. 5-33.

Ben-Akiva, M., Bierlaire, M. (2003) Discrete choice models with applications to departure time and route choice. *Handbook of transportation science*. Springer, pp. 7-37.

Ben-Akiva, M., Bolduc, D., Bradley, M. (1993) Estimation of travel choice models with randomly distributed values of time. *Transportation Research Record*, p. 88-97.

Ben-Akiva, M., Ramming, M.S. (1998) Lecture Notes: Discrete Choice Models of Traveler Behavior in Networks. Prepared for Advanced Methods for Planning and Management of Transportation Networks, Capri, Italy.

Ben-Akiva, M.E., Lerman, S.R. (1985) *Discrete choice analysis: theory and application to travel demand*. MIT press.

Benjamins, M., Lindveld, C., Van Nes, R. (2001) Multimodal travel choice modelling: a supernetwork approach. *Proceedings of In Proceedings 81st TRB Annual Meeting. Washington DC*, pp. Paper 03-2948 on CD-ROM.

Bhat, C.R. (1998) Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling. *Transportation Research Part A: Policy and Practice* 32, 495-507.

Bhat, C.R. (2000) Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling. *Transportation Science* 34, 228-238.

Bhat, C.R., Gossen, R. (2004) A mixed multinomial logit model analysis of weekend recreational episode type choice. *Transportation Research Part B: Methodological* 38, 767-787.

Bierlaire, M., Bolduc, D., McFadden, D. (2008) The estimation of generalized extreme value models from choice-based samples. *Transportation Research Part B: Methodological* 42, 381-394.

Bogers, E.A. (2009) *Traffic information and learning in day-to-day route choice*. Netherlands TRAIL Research School.

Bovy, P.H., Bekhor, S., Prato, C.G. (2008) The factor of revisited path size: Alternative derivation. *Transportation Research Record: Journal of the Transportation Research Board* 2076, 132-140.

Bovy, P.H., Stern, E. (1990) *Route Choice: Way Finding in Transport Networks*, Studie in Operational Reginal Science.

Bradley, M.A., Gunn, H.F. (1990) Stated preference analysis of values of travel time in the Netherlands. *Transportation Research Record*.

Brands, T., de Romph, E., Veitch, T., Cook, J. (2014) Modelling Public Transport Route Choice, with Multiple Access and Egress Modes. *Transportation Research Procedia* 1, 12-23.

Carlier, K., Fiorenzo-Catalano, S., Lindveld, C., Bovy, P. (2003) A supernetwork approach towards multimodal travel modeling. *Proceedings of Proceedings of the 82nd Transportation Research Board Annual Meeting, Washington DC*, pp. 10-2460.

Cascetta, E., Nuzzolo, A., Russo, F., Vitetta, A. (1996) A modified logit route choice model overcoming path overlapping problems: Specification and some calibration results for interurban networks. *Proceedings of Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, pp. 697-711.

Chapleau, R., Chu, K.K.A. (2007) Modeling transit travel patterns from location-stamped smart card data using a disaggregate approach. *Proceedings of 11th World Conference on Transport Research*.

Chen, L., Lv, M., Ye, Q., Chen, G., Woodward, J. (2011) A personal route prediction system based on trajectory data mining. *Information Sciences* 181, 1264-1284.

Cheong, C.C., Toh, R. (2010) Household interview surveys from 1997 to 2008–A decade of changing travel behaviours. *Editorial Team* 52.

Chowdhury, S., Ceder, A., Schwalger, B. (2015) The effects of travel time and cost savings on commuters' decision to travel on public transport routes involving transfers. *Journal of Transport Geography* 43, 151-159.

Chriqui, C., Robillard, P. (1975) Common bus lines. *Transportation Science* 9, 115-121.

Chu, C. (1989) A paired combinatorial logit model for travel demand analysis. *Proceedings of TRANSPORT POLICY, MANAGEMENT & TECHNOLOGY TOWARDS 2001: SELECTED PROCEEDINGS OF THE FIFTH WORLD CONFERENCE ON TRANSPORT RESEARCH*.

Clifton, K., Muhs, C. (2012) Capturing and representing multimodal trips in travel surveys: Review of the practice. *Transportation Research Record: Journal of the Transportation Research Board*, 74-83.

Daganzo, C.F., Sheffi, Y. (1977) On stochastic models of traffic assignment. *Transportation Science* 11, 253-274.

De Cea, J., Fernández, E. (1993) Transit assignment for congested public transport systems: an equilibrium model. *Transportation science* 27, 133-147.

de Grange, L., Raveau, S., González, F. (2012) A Fixed Point Route Choice Model for Transit Networks that Addresses Route Correlation. *Procedia - Social and Behavioral Sciences* 54, 1197-1204.

de la Barra, T., Perez, B., Anez, J. (1993) Multidimensional path search and assignment. *Proceedings of PTRC Summer Annual Meeting, 21st, 1993, University of Manchester, United Kingdom*.

Debrezion, G., Pels, E., Rietveld, P. (2009) Modelling the joint access mode and railway station choice. *Transportation Research Part E: logistics and transportation review* 45, 270-283.

Dial, R.B. (1967) Transit pathfinder algorithm. *Highway Research Record*.

Eluru, N., Chakour, V., El-Geneidy, A.M. (2012) Travel mode choice and transit route choice behavior in Montreal: insights from McGill University members commute patterns. *Public Transport* 4, 129-149.

EZlink (2013) EZLinK Card. pp. Introduction - EZlink Card.

Fiorenzo-Catalano, S., van Nes, R., Bovy, P.H. (2004) Choice set generation for multi-modal travel analysis. *European journal of transport and infrastructure research* 4, 195-209.

Florian, M. (2004) Finding Shortest Time-Dependent Paths in Schedule-Based Transit Networks: A Label Setting Algorithm. *Schedule-Based Dynamic Transit Modeling: theory and applications* eds Wilson, N.M., Nuzzolo, A. Springer US, pp. 43-52.

Flötteröd, G., Bierlaire, M. (2013) Metropolis–Hastings sampling of paths. *Transportation Research Part B: Methodological* 48, 53-66.

Fosgerau, M., Bierlaire, M. (2007) A practical test for the choice of mixing distribution in discrete choice models. *Transportation Research Part B: Methodological* 41, 784-794.

Fosgerau, M., Frejinger, E., Karlstrom, A. (2013) A link based network route choice model with unrestricted choice set. *Transportation Research Part B: Methodological* 56, 70-80.

Frejinger, E. (2008) Route choice analysis: data, models, algorithms and applications. École Polytechnique Federale de Lausanne.

Frejinger, E., Bierlaire, M. (2007) Capturing correlation with subnetworks in route choice models. *Transportation Research Part B: Methodological* 41, 363-378.

Frejinger, E., Bierlaire, M., Ben-Akiva, M. (2009) Sampling of alternatives for route choice modeling. *Transportation Research Part B: Methodological* 43, 984-994.

Friedrich, M., Hofsäß, I., Wekeck, S. (2001) Timetable-based transit assignment using branch and bound techniques. *Transportation Research Record: Journal of the Transportation Research Board* 1752, 100-107.

Greene, W.H., Hensher, D.A. (2003) A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological* 37, 681-698.

Grigolon, A.B., Borgers, A.W., Kemperman, A.D., Timmermans, H.J. (2014) Vacation length choice: A dynamic mixed multinomial logit model. *Tourism Management* 41, 158-167.

Guo, Z. (2011) Mind the map! The impact of transit maps on path choice in public transit. *Transportation Research Part A: Policy and Practice* 45, 625-639.

Guo, Z., Wilson, N.H. (2011) Assessing the cost of transfer inconvenience in public transport systems: A case study of the London Underground. *Transportation Research Part A: Policy and Practice* 45, 91-104.

Hamdouch, Y., Ho, H., Sumalee, A., Wang, G. (2011) Schedule-based transit assignment model with vehicle capacity and seat availability. *Transportation Research Part B: Methodological* 45, 1805-1830.

Hamdouch, Y., Lawphongpanich, S. (2008) Schedule-based transit assignment model with travel strategies and capacity constraints. *Transportation Research Part B: Methodological* 42, 663-684.

Han, B., Algers, S., Engelson, L. (2001) Accommodating drivers' taste variation and repeated choice correlation in route choice modeling by using the mixed logit model. *Proceedings of 80th Annual Meeting of the Transportation Research Board*.

Henn, V. (2000) Fuzzy route choice model for traffic assignment. *Fuzzy Sets and Systems* 116, 77-101.

Hensher, D.A., Greene, W.H. (2003) The mixed logit model: the state of practice. *Transportation* 30, 133-176.

Hess, S., Bierlaire, M., Polak, J.W. (2005) Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice* 39, 221-236.

Hess, S., Erath, A., Axhausen, K. (2008) Estimated value of savings in travel time in Switzerland: Analysis of pooled data. *Transportation Research Record: Journal of the Transportation Research Board*, 43-55.

Hess, S., Rose, J.M. (2009) Allowing for intra-respondent variations in coefficients estimated on repeated choice data. *Transportation Research Part B: Methodological* 43, 708-719.

Hess, S., Train, K.E. (2011) Recovery of inter-and intra-personal heterogeneity using mixed logit models. *Transportation Research Part B: Methodological* 45, 973-990.

Hofmann, M., Wilson, S.P., White, P. (2009) Automated identification of linked trips at trip level using electronic fare collection data. *Proceedings of Transportation Research Board 88th Annual Meeting*.

Hoogendoorn-Lanser, S. (2005) *Modelling travel behaviour in multi-modal networks*.

Hoogendoorn-Lanser, S., Bovy, P. (2007) Modeling overlap in multimodal route choice by including trip part-specific path size factors. *Transportation Research Record: Journal of the Transportation Research Board* 2003, 74-83.

Hoogendoorn-Lanser, S., Bovy, P., van Nes, R. (2007) Application of constrained enumeration approach to multimodal choice set generation. *Transportation Research Record: Journal of the Transportation Research Board* 2014, 50-57.

Hoogendoorn-Lanser, S., Bovy, P.H. (2004) Train Choice Behavior Modeling in Multi-modal Transport Networks. *Proceedings of Sustainable and Reliable Urban Accessibility Beijing Seminar 24/25 May 2004-05-16 Full papers*, p. 120.

Hoogendoorn-Lanser, S., van Nes, R., Bovy, P. (2005) Path size modeling in multimodal route choice analysis. *Transportation Research Record: Journal of the Transportation Research Board* 1921, 27-34.

Hurk, E.v.d., Kroon, L., Mar, G., x00F, ti, Vervest, P. (2015) Deduction of Passengers' Route Choices From Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems* 16, 430-440.

Jang, W. (2010) Travel time and transfer analysis using transit smart card data. *Transportation Research Record: Journal of the Transportation Research Board* 2144, 142-149.

Jánošíková, Ľ., Slavík, J., Koháni, M. (2014) Estimation of a route choice model for urban public transport using smart card data. *Transportation Planning and Technology* 37, 638-648.

Koppelman, F.S., Wen, C.-H. (2000) The paired combinatorial logit model: properties, estimation and application. *Transportation Research Part B: Methodological* 34, 75-89.

Kusakabe, T., Iryo, T., Asakura, Y. (2010) Estimation method for railway passengers' train choice behavior with smart card transaction data. *Transportation* 37, 731-749.

Lai, X., Bierlaire, M. (2015) Specification of the cross-nested logit model with sampling of alternatives for route choice models. *Transportation Research Part B: Methodological* 80, 220-234.

Lam, W.H., Zhou, J., Sheng, Z.-h. (2002) A capacity restraint transit assignment with elastic line frequency. *Transportation Research Part B: Methodological* 36, 919-938.

Leurent, F. (2008) Modelling seat congestion in transit assignment.

Leurent, F. (2009) On Seat Congestion, Passenger Comfort, and Route Choice in Urban Transit: Network Equilibrium Assignment Model with Application to Paris. *Proceedings of Transportation Research Board 88th Annual Meeting*.

Li, D., Miwa, T., Morikawa, T., Liu, P. (2016) Incorporating observed and unobserved heterogeneity in route choice analysis with sampled choice sets. *Transportation Research Part C: Emerging Technologies* 67, 31-46.

Louviere, J.J., Hensher, D.A., Swait, J.D. (2000) *Stated choice methods: analysis and applications*. Cambridge University Press.

Mai, T. (2016) A method of integrating correlation structures for a generalized recursive route choice model. *Transportation Research Part B: Methodological* 93, 146-161.

Mai, T., Fosgerau, M., Frejinger, E. (2015) A nested recursive logit model for route choice analysis. *Transportation Research Part B: Methodological* 75, 100-112.

Manski, C.F., Lerman, S.R. (1977) The Estimation of Choice Probabilities from Choice Based Samples. *Econometrica* 45, 1977-1988.

McFadden, D. (2000) Disaggregate behavioral travel demand's RUM side. *Travel Behaviour Research*, 17-63.

McFadden, D., Train, K. (2000) Mixed MNL models for discrete response. *Journal of applied Econometrics* 15, 447-470.

Ministry of Transport, S. (2010) Written Answer to Questions on Frequency and Capacity of MRT Trains. Singapore.

Nakayama, S., Kitamura, R. (2000) Route choice model with inductive learning. *Transportation Research Record: Journal of the Transportation Research Board*, 63-70.

Nassir, N., Hickman, M., Ma, Z.-L. (2015) Behavioural findings from observed transit route choice strategies in the farecard data of Brisbane. *Proceedings of Australasian Transport Research Forum (ATRF), 37th, 2015, Sydney, New South Wales, Australia*.

Nguyen, S., Pallottino, S. (1988) Equilibrium traffic assignment for large scale transit networks. *European journal of operational research* 37, 176-186.

Nguyen, S., Pallottino, S., Malucelli, F. (2001) A modeling framework for passenger assignment on a transport network with timetables. *Transportation Science* 35, 238-249.

Nielsen, O.A. (2000) A stochastic transit assignment model considering differences in passengers utility functions. *Transportation Research Part B: Methodological* 34, 377-402.

Nuzzolo, A., Crisalli, U., Rosati, L. (2012) A schedule-based assignment model with explicit capacity constraints for congested transit networks. *Transportation Research Part C: Emerging Technologies* 20, 16-33.

Nuzzolo, A., Russo, F. (1996) Stochastic assignment models for transit low frequency services: Some theoretical and operative aspects. *Advanced methods in transportation analysis*. Springer, pp. 321-339.

Nuzzolo, A., Russo, F., Crisalli, U. (2001) A doubly dynamic schedule-based assignment model for transit networks. *Transportation Science* 35, 268-285.

Ortúzar, J.D., Willumsen, L.G. (2001) *Modelling transport*. Wiley Chichester:.

Park, K., Bell, M., Kaparias, I., Bogenberger, K. (2007) Learning user preferences of route choice behaviour for adaptive route guidance. *IET Intelligent Transport Systems* 1, 159-166.

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, 557-568.

Prashker, J.N., Bekhor, S. (1998) Investigation of stochastic network loading procedures. *Transportation Research Record: Journal of the Transportation Research Board* 1645, 94-102.

Prato, C.G. (2009) Route choice modeling: past, present and future research directions. *Journal of Choice Modelling* 2, 65-100.

Prato, C.G., Bekhor, S. (2006) Applying branch-and-bound technique to route choice set generation. *Transportation Research Record: Journal of the Transportation Research Board* 1985, 19-28.

Prato, C.G., Bekhor, S. (2007) Modeling route choice behavior: How relevant is the composition of choice set? *Transportation Research Record: Journal of the Transportation Research Board* 2003, 64-73.

Pravinvongvuth, S., Chen, A. (2005) Adaptation of the paired combinatorial logit model to the route choice problem. *Transportmetrica* 1, 223-240.

PublicTransport@sg (2013) Fares & Ticketing.

Ramming, M.S. (2001) Network knowledge and route choice. Massachusetts Institute of Technology.

Raveau, S., Muñoz, J.C., De Grange, L. (2011) A topological route choice model for metro. *Transportation Research Part A: Policy and Practice* 45, 138-147.

Revelt, D., Train, K. (1998) Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of economics and statistics* 80, 647-657.

Revelt, D., Train, K. (2000) Customer-specific taste parameters and Mixed Logit: Households' choice of electricity supplier. *Department of Economics, UCB*.

Revelt, D.A. (1999) Three discrete choice random coefficients papers and one police-crime study. University of California, Berkeley.

Rieser-Schüssler, N., Balmer, M., Axhausen, K.W. (2012) Route choice sets for very high-resolution data. *Transportmetrica*, 1-21.

Rilett, L., Park, D. (2001) Incorporating uncertainty and multiple objectives in real-time route selection. *Journal of transportation engineering* 127, 531-539.

Robinson, S., Narayanan, B., Toh, N., Pereira, F. (2014) Methods for pre-processing smartcard data to improve data quality. *Transportation Research Part C: Emerging Technologies* 49, 43-58.

Report FHWA-RD-98-107. FHWA, US Department of Transportation, Washington, DC (1998) *Recommended Procedures for Chapter 13, Pedestrians, of the Highway Capacity Manual*.

Schmöcker, J.-D., Bell, M.G., Kurauchi, F. (2008) A quasi-dynamic capacity constrained frequency-based transit assignment model. *Transportation Research Part B: Methodological* 42, 925-945.

Schmöcker, J.-D., Fonzone, A., Shimamoto, H., Kurauchi, F., Bell, M.G.H. (2011) Frequency-based transit assignment considering seat capacities. *Transportation Research Part B: Methodological* 45, 392-408.

Schmöcker, J.-D., Shimamoto, H., Kurauchi, F. (2013) Generation and calibration of transit hyperpaths. *Transportation Research Part C: Emerging Technologies* 36, 406-418.

Seaborn, C., Attanucci, J., Wilson, N.H. (2009) Analyzing multimodal public transport journeys in London with smart card fare payment data. *Transportation Research Record: Journal of the Transportation Research Board* 2121, 55-62.

sgwiki (2013) Kawasaki Heavy Industries C151. p. Train type for north south line and east west line.

Siikamäki, J.V. (2001) Discrete choice experiments for valuing biodiversity conservation in Finland. University of California, Davis.

Singapore Land Transport Authority (2013) Household Interview Travel Survey 2012: Public Transport Mode Share Rises to 63%.

SMRT (2014a) Buses.

SMRT (2014b) Trains.

Spiess, H., Florian, M. (1989) Optimal strategies: a new assignment model for transit networks. *Transportation Research Part B: Methodological* 23, 83-102.

Sumalee, A., Tan, Z., Lam, W.H. (2009) Dynamic stochastic transit assignment with explicit seat allocation model. *Transportation Research Part B: Methodological* 43, 895-912.

Sumalee, A., Uchida, K., Lam, W.H. (2011) Stochastic multi-modal transport network under demand uncertainties and adverse weather condition. *Transportation Research Part C: Emerging Technologies* 19, 338-350.

Sun, Y., Xu, R. (2012) Rail transit travel time reliability and estimation of passenger route choice behavior: Analysis using automatic fare collection data. *Transportation Research Record: Journal of the Transportation Research Board*, 58-67.

Tan, R., Adnan, M., Lee, D.-H., Ben-Akiva, M.E. (2015) New Path Size Formulation in Path Size Logit for Route Choice Modeling in Public Transport Networks. *Transportation Research Record: Journal of the Transportation Research Board* 2538, 11-18.

The Straits Times (2014) Public transport rides up for ninth consecutive year The Straits Times.

Train, K. (2001) A comparison of hierarchical Bayes and maximum simulated likelihood for mixed logit. *Documento de Trabajo, Universidad de California, Berkeley*.

Train, K. (2009) *Discrete choice methods with simulation*. Cambridge university press.

Transit, S. (2014) Transport Services.

Trépanier, M., Morency, C., Agard, B. (2009) Calculation of transit performance measures using smartcard data. *Journal of Public Transportation* 12, 79-96.

Trepanier, M., Morency, C., Blanchette, C. (2009) Enhancing household travel surveys using smart card data. *Proceedings of Transportation Research Board 88th Annual Meeting*.

Tsukai, M., Okumura, M. (2003) Analysis of Inter-City Passenger's Route Choice Behavior On Non-Shortest-Time Routes. *Journal of the Eastern Asia Society for Transportation Studies* 5.

Uges, R., Hoogendoorn-Lanser, S., Bovy, P. (2002) Modeling Route Choice Behavior In Multimodal Transport Networks. Delft University Press, TRAIL Studies in Transportation Science.

Utsunomiya, M., Attanucci, J., Wilson, N. (2006) Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation Research Record: Journal of the Transportation Research Board* 1971, 119-126.

Van der Waard, J. (1988) The relative importance of public transport trip time attributes in route choice. *Proceedings of 16th PTRC Summer Annual Meeting*, Bath, United Kingdom.

Van der Waerden, P., Borgers, A., Timmermans, H. (2004) Choice Set Composition in the Context of Pedestrians' Route Choice Modeling.

Proceedings of 83rd Annual Meeting of the Transportation Research Board, Washington, DC.

Van der Zijpp, N., Fiorenzo Catalano, S. (2005) Path enumeration by finding the constrained K-shortest paths. *Transportation Research Part B: Methodological* 39, 545-563.

Vovsha, P. (1997) Application of cross-nested logit model to mode choice in Tel Aviv, Israel, metropolitan area. *Transportation Research Record: Journal of the Transportation Research Board* 1607, 6-15.

Vrtic, M., Axhausen, K.W. (2003) The impact of tilting trains in Switzerland: a route choice model of regional-and long distance public transport trips. *Proceedings of 82nd annual meeting of the transportation research board, Washington, DC*, pp. 11-15.

Wen, C.-H., Koppelman, F.S. (2001) The generalized nested logit model. *Transportation Research Part B: Methodological* 35, 627-641.

Williams, H.C. (1977) On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning A* 9, 285-344.

Yai, T., Iwakura, S., Morichi, S. (1997) Multinomial probit with structured covariance for route choice behavior. *Transportation Research Part B: Methodological* 31, 195-207.

Yamamoto, T., Kitamura, R., Fujii, J. (2002) Drivers' route choice behavior: analysis by Data mining algorithms. *Transportation Research Record: Journal of the Transportation Research Board*, 59-66.

Yang, H., Kitamura, R., Jovanis, P.P., Vaughn, K.M., Abdel-Aty, M.A. (1993) Exploration of route choice behavior with advanced traveler information using neural network concepts. *Transportation* 20, 199-223.

Yen, J.Y. (1971) Finding the k shortest loopless paths in a network. *management Science* 17, 712-716.

Zhang, Y., Lam, W.H., Sumalee, A., Lo, H.K., Tong, C. (2010) The multi-class schedule-based transit assignment model under network uncertainties. *Public Transport* 2, 69-86.