A novel modelling approach of integrated taxi and transit mode and route choice using city-scale emerging mobility data

Rakhi Manohar Mepparambath[a,b], Yong Sheng Soh[a], Vasundhara Jayaraman[a], Hong En Tan[a], Muhamad Azfar Ramli[a]

[a] *Institute of High Performance Computing (IHPC), Agency for Science Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore*
[b] *Corresponding author, email: rakhimm@ihpc.a-star.edu.sg*

Abstract

The modelling of mode and route choices of public transport passengers is an essential component of travel demand modelling and transportation planning. Traditionally, choice models are trained using data from revealed and stated preference surveys which are not only cost and time intensive but also suffer from bias due to limitations in sample size. In this article, we present a novel approach utilizing a combination of emerging data sources for non-private modes, namely public transit smart card data, taxi GPS trajectory data, and taxi trips transaction data in order to calibrate an integrated taxi and transit mode and route choice model. We apply our approach for a case study of the taxi and public transport system of Singapore using datasets obtained from the local transport authorities. We solve the first mile and last mile data gaps by overlaying the individual transport nodes against geospatial land use data in order to identify their actual origin and destination points of each journey. To model the behavioral inter-dependencies between the choice of taxi and transit options, we tested a two-level nested logit model and a cross-nested logit model. Along with the commonly included exploratory variables such as in-vehicle travel time, transfer time and number of transfers, our model also incorporates differences in travel cost and separate mode specific constants for peak and off-peak periods. Generic and mode-specific in-vehicle travel time and cost coefficients are tested in the utility functions for the transit and taxi alternatives. Willingness to pay estimates are calculated and compared against similar estimates from Singapore. We also present an application of the model by predicting the mode split between taxi and transit under selected transit and taxi fare change scenarios. Our modelling methodology is highly generalizable and can be applied to other cities with similar data availability.

Keywords: Nested logit, cross-nested logit, mode choice, transit route choice, smart card data, taxi data

1. Introduction

Modelling the choices made by individual passengers such as the choice of mode and public transit route are inevitable steps in travel demand modelling. Traditionally, data from surveys are used in the development of dis-aggregate mode and transit route choice models. Surveys are cost intensive and survey data often has limited sample size and sampling bias. Smart card data has many advantages over surveys when used in the modelling of transit route choice behaviour of passengers. Unlike surveys, smart card data provides large volume of data that is passively collected over a long period of time without any additional effort on data collection. Depending on the type of Automated Fare Collection (AFC) system used, the smart card data can provide accurate information on the transit routes taken by the passengers, compared to traditional travel diaries

© <2023>. This manuscript version is made available under the CC-BY-NC-ND 4.0 license
https://creativecommons.org/licenses/by-nc-nd/4.0/

where missing information is a common occurrence and getting a good representative sample for the population can be challenging. In case of cities where smart card usage is predominant, this data set provides the public transit trip information for the majority of the public transport users. As a result, emerging data sources like transit smart card data are increasingly used by researchers in travel demand modelling applications.

Like the smart card data, researchers are showing growing interest in using taxi data sets like taxi GPS trajectory data and taxi trips transaction data for travel demand modelling purposes. Similar to transit smart card data, taxi data is also automatically collected and is available in large volumes for longer periods of time. Taxi data provides rich information on the taxi journeys such as the pick-up and drop-off locations, routes taken, travel time and taxi fare. Taxi data when used in combination with smart card data can be used to model the mode choice behaviour of passengers between taxi and other public transit options. Although the taxi data set is used in some travel demand modelling applications (Liu et al. (2015), Tang et al. (2015), Liu et al. (2012), Zhan et al. (2013), Li et al. (2017)), it is not used in combination with the smart card data to model the mode and transit route choice behaviour of passengers. Considering taxis along with other public transit modes are especially important for cities that has non-negligible mode share for taxis. For example, in Singapore 67% of peak hour mode share in 2016 was contributed by non-private modes which includes 5% of taxi, 34% of bus and 28% of train. Hence, by including taxis along with other public transit modes, we will have a more holistic view of the passenger choice behaviour among the non-private modes of transport.

This study applies a novel data fusion approach to combine two automatically collected large scale data sets—transit smart card data and taxi data sets—to develop a nested mode and transit route choice model. The estimated model from this study can be used in predicting the travel demand for non-private modes and also in evaluating how potential changes to public transportation can influence travel mode choice and transit route choice.

2. Literature review

In this section we review the relevant literature on smart card and taxi data and their applications to mode and transit route choice models.

Smart card data has been used in a variety of applications within the broad umbrella of travel demand modelling such as transit o-d matrix estimation, analysing aggregate or individual travel patterns, and transit route choice modelling. Hussain et al. (2021) presented a critical review of o-d matrix estimation studies that used smart card data. o-d matrix is an essential input for transit network analysis and smart card data with boarding only information was used in many studies (Munizaga and Palma (2012), Kang et al. (2021), Kumar et al. (2018)) to estimate the transit o-d matrix. Transfer detection algorithms that distinguish a transfer from an activity between two consecutive validated stops is an important feature of such studies. Some studies (Alsger et al. (2015), Cheng et al. (2020)) also used smart card data with boarding and alighting information to test or validate the o-d matrix estimation algorithms.

Another stream of studies used smart card data to analyze the aggregate travel patterns within a city (Ma et al. (2013), Kieu et al. (2015), Lei et al. (2020), Liu et al. (2020), Zhong et al. (2015)). Clustering algorithms were typically used in these studies to identify the spatial and temporal patterns of travel. Individual mobility patterns were also identified from smart card data using data mining techniques and data-driven models (Ma et al. (2017), Zhao et al. (2018), Foell et al. (2014), Goulet-Langlois et al. (2017), Lathia et al. (2013)).

More recently, smart card data is also used in transit route choice modelling. Earlier transit route choice models were developed using primary data that is collected either using a revealed preference survey (Eluru et al. (2012), Hunt (1990), Lo et al. (2004), Raveau et al. (2011), Anderson et al. (2017)) or a stated preference survey (Kurauchi et al. (2012)). Combined revealed and stated preference survey data was also used in some studies (Pursula and Weurlander (1999), Lam and Xie (2002)) to estimate transit route choice models. As more and more transit systems use smart card based AFC system, there is a growing interest on using smart card data in transit route choice modelling (Jánošíková et al. (2014), Van Der Hurk et al. (2014), Zhao et al. (2016), Nassir et al. (2019), Kim et al. (2020)). Multinomial logit (Hunt (1990), Raveau et al. (2011), Pursula and Weurlander (1999), Jánošíková et al. (2014), Nassir et al. (2019)), path-size logit (Lam and Xie (2002), Anderson et al. (2017) Yap et al. (2020)) and nested logit (Lo et al. (2004)) were the most common choice models used in transit route choice modelling. All the previous studies modelled the choices within the transit options without considering the influence of available non-private mode like taxi on these choices.

Separately, taxi data has been used in studying the mobility patterns (Liu et al. (2015), Tang et al. (2015), Liu et al. (2012)), urban forms (Liu et al. (2015), Liu et al. (2012)), and network link travel speeds (Zhan et al. (2013)). Previous studies that used transit and taxi data together include Zhu et al. (2016), where the recovery patterns for roadway and subway systems of New York City from Hurricanes Irene and Sandy was studied using taxi data and subway turnstile ridership data. Another study that used transit smart card data and taxi data is Li et al. (2017), where they examined the spatial variation of urban taxi ridership due to the impact of a new subway line by analysing the taxi and transit data. To the best of our knowledge, no other studies have used the approach used in this study where we combine these two data sets to estimate a joint mode and transit route choice model.

To develop the joint model for mode and transit route choice, we combine the smart card data with taxi data using an o-d assignment procedure that consider the land use type and density of buildings in the vicinity of observed journey start and end points. In order to combine the observed demand from the transit smart card data and taxi data, passenger journeys needs to be assigned to the same set of origin-destination (o-d) pairs. The problem with using two disparate data sources from two different modes of transport is that the spatial resolution of the journey start and end points differ; journeys arising from smart card data begins from bus stops and train stations, while taxi journeys originate and end at road network nodes and edges. We therefore generate a new common travel demand node set from which the passenger journeys start and end. The journeys from smart card data and taxi data are assigned to these common nodes based on weighted assignment using land use building weightage data.

Using the combined data set, we calibrate a nested logit model and a cross-nested logit model that jointly model the mode and public transit route choice behaviour of passengers. To develop the choice model, we also developed models to generate 1) the choice set of alternate public transit routes and their attributes and 2) attributes of taxi journeys for any given o-d pair. Most of the existing work model mode and transit route choice separately, whereas the mode choice—especially the choice of non-private mode like taxi—can highly be influenced by the available transit options. For example, Li et al. (2017) found that opening of a new subway line decreases the taxi trips volume by 40% - 50%, if both the origin and destination of the taxi trips were within 1 km from the subway station. In this study, we explicitly model this behavioural inter-dependencies by considering a nested structure where the first level of choice is between the taxi and transit options and within the transit nest we consider the available multi-modal transit

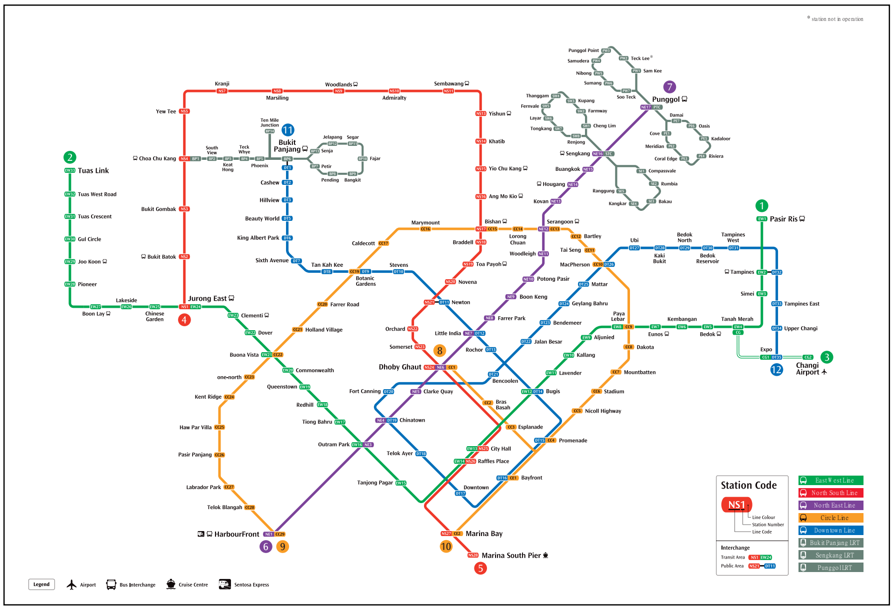


Figure 1: Layout of the Singapore MRT and LRT network in 2018

routes. One of the main objectives of this study is to model the mode and transit route choice behaviour using emerging revealed preference data sources like transit smart card data and taxi data. Our methodology also demonstrates how the data gaps in combining smart card data and taxi data can be addressed using other available data sets. The proposed methodology is demonstrated using data sets from Singapore, but is applicable in cities where similar data sets are available.

3. Study area and Data description

Study area

Singapore is a city state with a land area of 721.5 square kilometer and a population of 5.6 million. Due to it's limited land space and high density of population, Singapore has invested heavily on building a vast public transport system that includes heavy rail (Mass Rapid Transit or MRT), light rail (Light Rail Transit or LRT) and bus modes. The MRT and LRT network from Singapore is shown in Figure 1. Singapore MRT network has more than 130 stations across six MRT lines that span across the island with 200km MRT lines and over three million daily ridership. The LRT network consists of more than 40 stations across two LRT lines with 28 km LRT lines and over 200 thousand daily ridership. Singapore also has a vast bus network with more than 5000 bus stops and more than 500 bus services with over four million daily ridership. Singapore has a controlled private car ownership and taxis, a personalized form of public transport, fills this gap by supplementing the train and bus modes. Along with transit smart card and taxi data sets, transit network and land-use data sets are also used in our study to fill the data gaps in the first two data sets. The four different data sets from Singapore that are used in this study are briefly described below.

Transit smart card data

EZlink, a contactless, stored value smart card is the dominant method of public transport fare payment in Singapore. Payment for all public transport trips can be made using EZlink cards regardless of the mode (MRT, LRT and bus) or the operator. Unlike many other Automatic Fare Card systems that use flat fare scheme for public transport trips, AFC system in Singapore uses a distance-based fare scheme, where passengers have to tap in at the boarding point and tap out at the alighting point. As a result, the EZlink data set contains detailed information on each transit trip that include the mode, boarding, alighting stops and corresponding time stamps. Each EZlink card has a unique card ID and consecutive trips made by the same card that satisfy the transfer rules are considered as a single journey. Hence each journey in EZlink data is assigned a unique journey ID which may contain more than one trip, the order of which is distinguished by the transfer number provided in the data set. EZlink data from a single weekday of Feb 2018 is used in this study to identify the chosen transit route for a given journey and it has roughly 6.2 million trips, which when combined using transfer information, translates into 4.5 million journeys.

Taxi data

The taxi data set comprises of two separate data sets : 1) taxi GPS trajectory data and 2) taxi trips transaction data. Taxi data from the same time period as EZlink data (Feb 2018) is used in our analysis. The taxi GPS trajectory data contains taxi GPS traces of all taxis in Singapore for the month of Feb 2018. This data set includes the location information for approximately 21000 unique taxis in Singapore on an average weekday. The taxi trips transaction data holds information on individual passenger trips from their origin to destination. Approximately 420000 taxi trips are recorded for an average weekday in Feb 2018. Each trip record contains information of taxi ID, pick-up time, pick-up location, drop-off time, drop-off location, trip distance and the taxi fare charged for that trip. The taxi data is used in this study to: 1) identify the choice of taxi mode for a journey between a given o-d, and 2) develop models that predict the taxi attributes between any given o-d pair.

Transit network data

The transit network data comprises two components: 1) the bus network data, and 2) the train network data. The bus network data contain information on: i) the location of all bus stops, and ii) the sequence of stops visited by every bus service. In addition, the data set provides the (estimated) timings of the first and the last buses for every service within a single day as well as the expected bus headways, which is further differentiated into morning/afternoon and peak/off-peak hour timings. These data sets are provided and maintained by the Land Transport Authority of Singapore and are publicly available from the DataMall website (LTA (2021)). The train network data comprises: i) the location of all MRT and LRT stations as well as ii) the connections between stations. In addition, we also use the first/last train timing schedule, which are published on the train operator websites. . These datasets are used in this study to generate the choice set of alternative transit routes and attributes of these routes between any given o-d pair.

Land-use data

The land use dataset used in this study is the Urban Redevelopment Authority of Singapore's (URA) MasterPlan 2014 Land-use dataset (Figure 2). This data set consists of approximately 111,000 two-dimensional polygons that divides the entire land area of Singapore into disparate land parcels. The polygons contain two key pieces of information utilised in our modelling: 1) the categorisation of land-use (residential, commercial, industrial, mixed use, transport infrastructure and other miscellaneous classifications such as waterbodies and green space), and 2) the gross plot ratio (GPR) of the land parcel. The GPR value provides a guide for the maximum allowable amount of urban floor space (which can be much larger than actual land area due to possibility of having multiple building floors) as a ratio of the area of the land parcel and is a value that

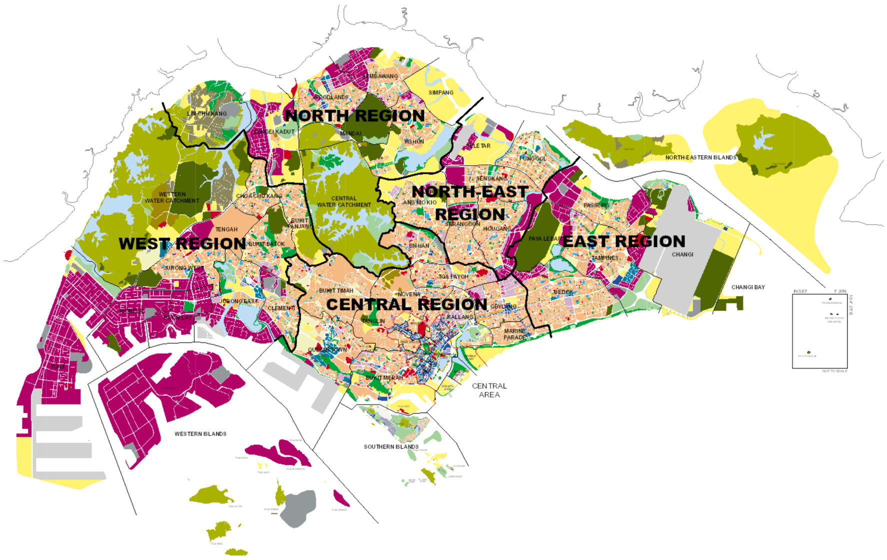


Figure 2: Singapore's Land Use Master Plan 2014 (source: URA (2014))

ranges from 1.0 to 15.0, which we use as the weight when assigning passenger flows to that land parcel.

4. Methodology

4.1. Sampling

The mode choice model discussed in this study excludes private modes and is applicable only to the passengers who do not have access to private modes. Hence, we first try to identify the passengers in the transit smart card data who do not have access to private modes and sample only the transit journeys made by such passengers for the model estimation. Following procedure is used to identify passengers who are captive to non-private modes. The smart card data has a unique identifier for each card, which we assume as a unique passenger. We analyzed the smart card data for the whole month of February 2018 and found the frequency of travel by each passenger over weekdays and weekends. A passenger is identified as a captive passenger to non-private modes if his or her travel frequency falls above 50 percentile of the passenger travel frequency during weekdays and above 25 percentile during weekends. This is based on the assumption that a passenger who makes regular transit trips during weekdays and who also makes a reasonable number of transit trips during weekends are less likely to have access to private modes.

At the same time, it is not possible to identify such passengers from taxi data as there are no identifier for the person making the trip. To avoid the sampling bias due to this, this study use the aggregate mode shares between transit and taxi modes for individuals without private modes as the sampling proportions for transit and taxi trips. The aggregate mode shares are derived from the Household Interview Travel Survey (HITS) data. HITS is a large scale travel behavioural survey conducted by Land Transport Authority (LTA) of Singapore at every four to five years. The survey collects information on socio-demographic details of the respondent, his or her household details along with their travel details. Since the sampling methodology used in the current study is choice-based and we explore different classes of discrete choice models including models that fall under Generalized Extreme Value models, we also use the weighted conditional maximum likelihood (WCML) proposed by Bierlaire et al. (2008) while estimating the parameters of the choice model to avoid the potential sampling bias.

4.2. Data fusion and travel demand generation

Before public transport and taxi data sets can be combined together, two main processing steps are required. Firstly, individual trips in the smart card data are chained into a multi-modal journey based on transfer rules used by the transit fare system in Singapore: a public transport journey is a series of single-mode trips with up to five transfers within a two hour time window, and a maximum of 45 minutes allowed for each transfer between bus and rail, or between different bus services.

Secondly, since public transport journeys and taxi trips have different geographical resolution in terms of their origins and destinations, there is a need to define a set of origin and destination locations that is common to passengers utilising these two modes. If we assume that passengers ultimately originate and complete their commute in buildings, the polygonal objects that represent such buildings in the land-use data set can be used as these common travel demand generation nodes.

We first perform filtering to remove polygons classified as green reserves, water bodies, roads, rail lines and other transportation infrastructure as they do not provide relevant origin and destination locations for urban commuting. The initial filtering results in about 100 thousand polygons remaining in the data set but with highly non-uniform spacings and size of polygons. In order to reduce the data set down to a set that is computationally feasible for passenger generation as well as to make the distribution of polygons more uniform, we employ the DBSCAN clustering algorithm (Ester et al. (1996)) on the centroids of each polygon and combine the smaller clustered land parcels together. To retain the original GPR information, each newly formed clustered polygon is assigned the average of the original polygon GPR values, weighted by the size of each polygon and this value now represents the land use weight (w_o) of each clustered polygon. The set of centroids ($\tilde{C}$) of these clustered polygons are now suitable to serve as the common travel demand node set for origin and destination locations for our entire study area.

In order to assign the observed travel demand from each of the origin (O) and destination (D) transport nodes to these land use parcels, we use the following probabilistic assignment model (schematic shown in Figure 3).

1. From the full travel demand node set $\tilde{C}$, identify the subset of all origin nodes $\tilde{o} = \{o_1, o_2, o_3, ..\}$ and destination nodes $\tilde{d} = \{d_1, d_2, d_3, ..\}$ within a specified threshold distance r_t around each transport node, i.e. $\tilde{o} = \{c \in \tilde{C} \,|\, dist(c, O) \leq r_t\}$ & $\tilde{d} = \{c \in \tilde{C} \,|\, dist(c, D) \leq r_t\}$.
2. Using the total commuting demand $f_{O \to D}$ from the transport node origin O and destination D, we assign proportionally the flow $f_{o \to d}$ between the neighboring origin o and destination d nodes using the following formulation:

$$f_{o \to d} = \frac{w_o}{\sum\limits_{o \in \tilde{o}} w_o} \frac{w_d}{\sum\limits_{d \in \tilde{d}} w_d} f_{O \to D} \tag{1}$$

We assume that passengers stick to walking for the first mile and last mile travel between the travel demand node and the transport nodes and the threshold distance r_t is fixed based on the walkable distance between the travel demand node and the transport node.

4.3. Public transport route set and attribute generation

In order to estimate the mode and public transit route choice model, the choice set of alternative public transit routes and their attributes need to be known for any given o-d in the data that is used in the choice model estimation. For example, between a

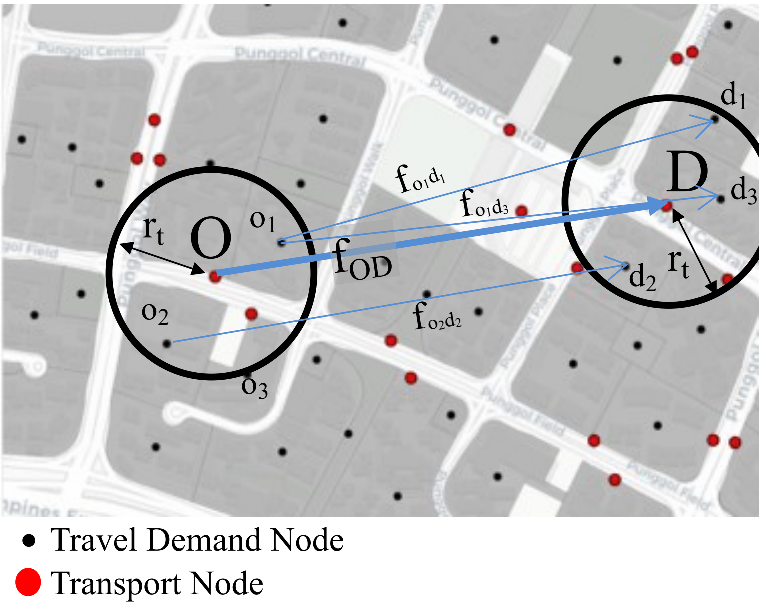


Figure 3: Illustration of the probabilistic demand assignment model from transport nodes (red circles) to the common travel demand nodes (black circles)

given origin A and destination B, let us assume that there are multiple routes r_1, r_2, r_3 and r_4 that are available for a passenger. An entry in the smart card data for the origin-destination pair $A - B$ will show that a passenger took one route, say r_1 among all possible routes available in the route choice set, $R = \{r_1, r_2, r_3, r_4\}$. The smart card data can then provide the attributes of the observed route r_1, but for the choice model estimation, the full choice set $R = \{r_1, r_2, r_3, r_4\}$ and the attributes of all the alternatives r_1, r_2, r_3 and r_4 in the choice set are needed.

Two approaches of choice set generation are found in studies that use smart card data in Ftransit route choice model estimation. In the first approach, the choice set is directly deduced using all observed routes between an o-d pair in the smart card data (Kim et al. (2020), Jánošíková et al. (2014), Raveau et al. (2014)). In this approach, only o-d pairs with multiple observed routes in the smart card data can be used for the choice model estimation. In dense transit networks passengers may be taking only one particular route between an o-d not because of non-availability of alternative routes, but because of dominance of one route over others. Hence, discarding such o-d pairs completely from the model estimation may not be the best approach. Another disadvantage of this approach is that the application of the choice model in predicting the route is limited to such observed o-d pairs from the smart card data.

The second approach in choice set generation is to make use of the transit network data to generate a list of possible routes between every o-d pair (Anderson et al. (2017)) – this is the approach that we take. Using the network data, we first built the transit network that include all nodes that are bus stops or train stations and all links that are existing bus routes between the bus stops or train link between the train stations. All possible routes between the o-d pairs in this network are first generated. However, the number of alternatives is frequently very large between a given o-d pair, of which many of them are implausible and would never be considered by passengers, hence it is necessary to generate this list strategically. We use the approach explained below to reduce the number of alternative routes in the choice set.

In our implementation, we list down all possible routes of the following types: (i) Bus, (ii) Bus-Bus, (iii) Rail, (iv) Bus-Rail, (v) Rail-Bus, (vi) Bus-Rail-Bus. In computing these routes, passengers are permitted to make walking transfers up to a distance of 400 metres if the transfers do not involve a rail station, and up to a distance of 1000 metres if these transfers involve a rail station. Routes are always identified by the sequence of services. As such, in instances where it is possible to take the same sequence of services, for example bus service B_1 followed by bus service B_2 between an o-d pair but make the transfer between B_1 and B_2 at different bus stops due to the overlap in bus routes of these services, these route choices will be identified as being equivalent. Our procedure takes into account the start time and the end time of each bus service or train service; if a passenger arrives at a bus stop or rail station at a timing where certain services are not available, these services are excluded.

The list of all possible routes between o-d pairs is frequently very large and hence it is necessary to employ certain heuristics to speed up the computation process. In our implementation, we only retain the fastest five routes in each of the six categories (Bus, Bus-Bus, etc.). We note that it is important that we treat these categories separately as some passengers may choose routes having fewer changes even if they take longer compared to other alternatives.

The route choice model requires as input a series of attributes of each route in the choice set. The journey distance for a route is obtained by summing the distance of every transit trip in the route, rounded to the nearest 0.1km. The distances for bus trips are available from DataMall (LTA (2021)) and the distances for rail trips are publicly available. The total walking distance is obtained by summing the distance of all estimated walking transfers. This figure is not included in the journey distance. The estimated journey timings are inferred using the first/last bus schedule timings from DataMall (LTA (2021)) and the first/last train timings available from the train operator website. The transit fare for each journey is calculated using the fare structure set by Singapore Public Transport Council (PTC). The fare structure is based on the journey distance, and prices bus and rail trips equally. The details of the transit fare structure can be found in PTC (2019). Finally, we estimate all walking transfers to take 1.2 seconds per map-distance metre.

4.4. Taxi attribute generation

We propose that taxi travel time and taxi fare are two important taxi attributes that influence passengers' mode choice behaviour. While estimating mode and transit route choice model, we will need taxi attributes for all journeys, including those who took the transit options. Hence, we use the taxi data to develop models that predict the taxi travel time and taxi fare for any given o-d.

Each record in the taxi GPS trajectory data set contains a unique ID to identify a taxi and its geo-position (latitude, longitude) information for the recorded time, with a time resolution of about 30 seconds. It is therefore possible to estimate travelling distance and speed of the taxi from the consecutive records in the data set. Since taxi GPS trajectory data set contains errors in the form of false or erroneous pings owing to system faults, we remove all location records that are outside Singapore boundary as a first step. Because the time resolution of the taxi GPS trajectory data set is 30 seconds, we can run a standard map-matching algorithm to map the taxi trajectories to our underlying road network. The map-matching algorithm further removes erroneous pings that show the taxi moving impossibly longer distances than realistically achievable.

From the map-matched taxi GPS trajectory data set obtained as described above, we can extract distance travelled and speed of each taxi along each road in our underlying road network at any given time during any day of the month of Feb 2018. For the sake of simplicity, we divide each 24 hour day into 15 minute time windows, and determine the median speed observed from the map-matched taxi GPS trajectory data set along each road in the road network over 96 time windows in a 24 hour period. This information is then used to build a matrix containing travel time along each road for all 96 time windows within a 24 hour period. Gaps in the matrix due to any road not experiencing taxi traffic can be filled using minimum travel time along the road calculated using minimum speed and length of road. We use information in publicly available speed-band data to fill in these gaps in our matrix.

The taxi trips transaction data set gives us information on the number of passenger trips by taxi, the start and end times of the trip, origin and destination locations of each trip, and trip fare for the same period as the taxi GPS trajectory data set. The taxi trip origin and destination locations are matched to the closest road network node, and a shortest path between the origin and destination road nodes is computed using Contraction Hierarchy as described in Geisberger et al. (2012) using the information in the travel time matrix as edge weights in the road network. Thus for any pair of origin and destination locations, we are able to determine travelling paths, travel time and travel distance. We also develop a taxi fare model, which is a simple linear regression model with travel distance as the explanatory variable and is fitted using the taxi trips transaction data.

4.5. Mode and transit route choice model

We use a Random Utility Maximization (RUM) model to represent the passengers' mode and transit route choice behaviour. In a choice model based on the random utility theory, a decision maker chooses an alternative that maximizes his or her utility. As the utility perceived by the decision maker is not observable to the modeller, utility is assumed to be a random variable with two parts. The first part is the observable/deterministic part which is expressed as a function of observable variables. The parameters of this function are estimated using data. The second part of the utility is the unobservable part, which is expressed as a random error term. To derive choice probabilities, distributional assumptions are made on this random error term and parameters of the distribution are fitted using data.

In our case, we have a multidimensional choice problem where passengers choose between the mode (transit vs taxi) first and the route in the transit (bus-train) network next. We use a nested logit structure as shown in Figure 4 to represent this multidimensional choice problem. In a nested logit structure, the correlation between alternatives in the same nest is explicitly considered. This assumption is true in our case, where different transit routes between an o-d pair can be correlated. Nested logit structure is used in route choice modelling to overcome the route overlap issue (Vovsha and Bekhor (1998)). Moreover, the nested logit model does not have the most limiting Independence from Irrelevant Alternatives (IIA) property of a standard logit model. The exploratory variables included in the utility function are selected based on the existing literature and the availability of the variable in the data set. In-vehicle travel time, walking time for transfers, number of transfers and fare are considered in the model as exploratory variables. The transit nest, denoted as PT have n routes and the utility for each route $r = r_1, r_2, ...r_n$ is shown in Equation 2.

$$
\begin{aligned}
U_{PT_{r_1}} = \beta_{IVTT} \times IVTT_{PT_{r_1}} + \beta_{WT} \times WT_{PT_{r_1}} + \beta_{NoT} \times NoT_{PT_{r_1}} + \beta_{Fare} \times Fare_{PT_{r_1}} \\
+\beta_{PT} + \epsilon_{PT_{r_1}} \\
. \\
. \\
U_{PT_{rn}} = \beta_{IVTT} \times IVTT_{PT_{rn}} + \beta_{WT} \times WT_{PT_{rn}} + \beta_{NoT} \times NoT_{PT_{rn}} + \beta_{Fare} \times Fare_{PT_{rn}} \\
+\beta_{PT} + \epsilon_{PT_{rn}}
\end{aligned}
\tag{2}
$$

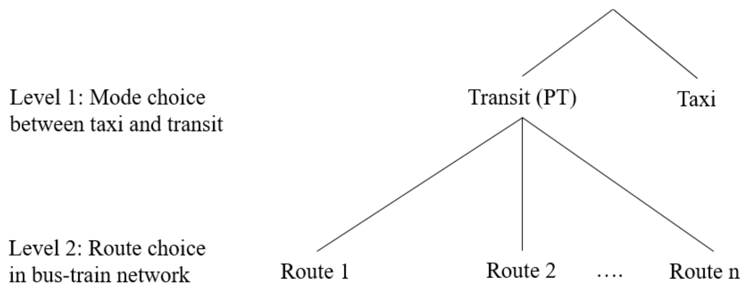


Figure 4: The two-level nested logit structure for mode and transit route choice mode
l

The utility function for the taxi nest which has only one alternative is presented in Equation 3.

$$U_{Taxi} = \beta_{IVTT} \times IVTT_{Taxi} + \beta_{Fare} \times Fare_{Taxi} + \beta_{Taxi} + \epsilon_{Taxi} \tag{3}$$

Where $IVTT$ = In-vehicle travel time, WT = Walking time for transfers, NoT = Number of transfers and $Fare$ are the independent variables considered in the model. β corresponds to the respective parameter of the utility function and ϵ is the random error term, which is assumed to be Gumbel distributed with a scale parameter μ_{PT} for the PT nest and μ_{Taxi} for the Taxi nest. The choice probabilities in a nested logit model can be calculated using the following equations. The marginal probability of choosing Transit(PT) or Taxi is given by Equation 4.

$$\begin{aligned} P_{PT} &= \frac{e^{V_{PT}+\frac{I_{PT}}{\mu_{PT}}}}{\sum_{m\in M} e^{V_m+\frac{I_m}{\mu_m}}} \\ P_{Taxi} &= \frac{e^{V_{Taxi}}}{\sum_{m\in M} e^{V_m+\frac{I_m}{\mu_m}}} \end{aligned} \tag{4}$$

Where $M = \{PT, Taxi\}$ is the set of all modes, $I_{PT} = ln \sum_{r\in R_{PT}} e^{\mu_{PT}V_r}$, is the log sum term or inclusive utility of transit nest and $R_{PT} = \{r_1, r_2, ..., r_n\}$ is the set of all transit routes for an o-d. The conditional probability of choosing route r_1 given transit is chosen is given by Equation 5 and the marginal probability of choosing route r_1 is given by Equation 6.

$$P_{r_1|PT} = \frac{e^{\mu_{PT}V_{r_1}}}{\sum_{r\in R_{PT}} e^{\mu_{PT}V_r}} \tag{5}$$

$$P_{r_1} = P_{r_1|PT} P_{PT} \tag{6}$$

The two-level nested logit model discussed above does not capture the possible correlation among different types of transit routes within the transit nest. Also transit

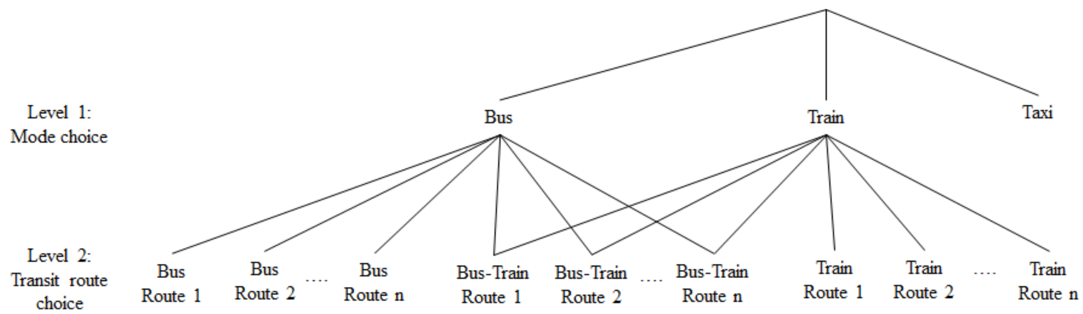


Figure 5: Cross-nested logit structure for combined mode and transit route choice

and taxi alternatives are likely to have different willingness to pay values, which is not considered in the utility functions with generic parameters given in Equation 2 and 3. To capture this, we also tested a cross-nested logit model where the transit routes in the PT nest are grouped based on the mode and separate $IVTT$ and $Fare$ coefficients for transit (β_{IVTT}^{PT}, β_{Fare}^{PT}) and taxi (β_{IVTT}^{Taxi}, β_{Fare}^{Taxi}) alternatives are considered. The cross-nested logit model structure is shown in Figure 5. All transit routes with only bus mode are grouped under the bus nest and those routes with only train mode are grouped under the train nest. Routes with bus and train modes are included in both bus and train nests with nest membership parameters denoted by α_{Bus} and α_{Train}, where $0 \leq \alpha_{Bus} \leq 1$, $0 \leq \alpha_{Train} \leq 1$ and $\alpha_{Bus} + \alpha_{Train} = 1$. The conditional probability of choosing a bus-train route BTr_1 given the bus nest is selected is given by Equation 7. Here the summation in the denominator is over all routes in the bus nest (R_B) that include bus only routes and bus-train routes. As bus only routes only belong to the bus nest α_{Bus} values for such routes are set to one. The probability of choosing bus mode in a journey is given by Equation 8. and the marginal probability of choosing a bus-train route is given by Equation 9.

$$P_{BTr_1|Bus} = \frac{\alpha_{Bus}^{\mu_{Bus}} e^{\mu_{Bus} V_{BTr_1}}}{\sum_{r \in R_B} \alpha_{Bus}^{\mu_{Bus}} e^{\mu_{Bus} V_r}} \tag{7}$$

$$P_{Bus} = \frac{\left(\sum_{r \in \{R_B\}} \alpha_{Bus}^{\mu_{Bus}} e^{\mu_{Bus} V_r}\right)^{\frac{1}{\mu_{Bus}}}}{\sum_{m \in M} \left(\sum_{r \in R} \alpha_m^{\mu_m} e^{\mu_m V_r}\right)^{\frac{1}{\mu_m}}} \tag{8}$$

$$P_{BTr_1} = \sum_{m \in M} P_{BTr_1|m} P_m \tag{9}$$

The parameters of the nested and cross-nested logit models are estimated using the weighted conditional maximum likelihood (WCML) estimator available in PandasBiogeme software (Bierlaire (2018)).

5. Results

5.1. Travel demand generation results

To understand the travel demand patterns, we analysed the spatial and temporal distributions of the observed demand from the transit and taxi options. The temporal

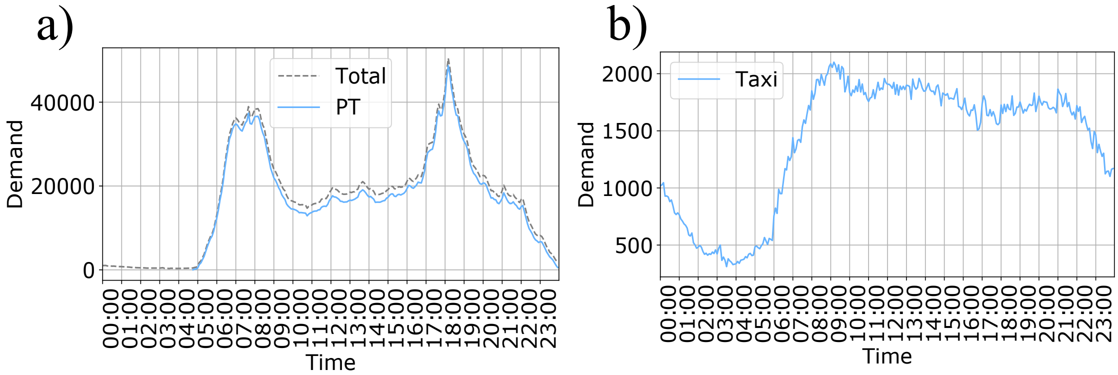


Figure 6: Temporal distribution of total observed travel demand ((a), dashed grey line), Public Transport ((a), blue line) and Taxi (b) respectively across a 24 hour period for a single weekday in Singapore in Feb 2018.

distribution of the total demand and demand by taxi and transit modes are shown in Figure 6. The public transport demand clearly has the morning and evening peaks, whereas the demand for taxi does not show much variations across different hours of the day, except for decrease in demand during early morning hours.

The spatial distribution of the total demand over AM peak period (0630h - 0830h) and PM peak periods (1700h - 1900h) are shown in Figure 7. The destination with highest demand during the AM peak and the origin with highest demand during the PM peak is found to be the same and is located in the CBD area of Singapore that has high commercial land-use, revealing the commuting pattern to and from the CBD. The origins with higher demand during the AM peak and the destinations with higher demand during the PM peak are more distributed across the island. These locations roughly coincides with the heavily populated residential areas in Singapore (see Figure 2 for Singapore land-use plan). Hence, the spatial distribution of demand during AM and PM peak hours clearly show the morning and evening commuting behaviour of the passengers.

5.2. Public transport choice set generation results

As we noted in Section 4.3, it is necessary to restrict the total number of route choices available to each passenger for computational reasons. However, this resulted in a problem that in many instances the observed journey is not among the route choice set provided by the algorithm. While undesirable, the situation is to be expected because (i) passengers may choose routes based on reasons beyond those that our model accounts for, and (ii) passengers are simply not aware of the optimal set of route choices that are available to them. In order to overcome this problem, we adopt the procedure found in Anderson et al. (2017) where an observed route if not already in the generated choice set is added to the choice set to have the final choice set that is used in the choice model estimation. Subsequently, we obtain 100% coverage of observed routes in the choice set.

We also compare the transit route attributes generated by the model with the corresponding attribute from the smart card data. Only transit in-vehicle travel time and transit distance are available directly in the smart card data for the comparison and hence we restrict our comparison to these two transit attributes. The comparison results are shown in Figure 8, where the Pearson correlation coefficient, r and Root Mean Square Error, $rmse$ between the observed and predicted transit attributes are also shown. The comparison results shows reasonably good performance for our application, where we use these attributes as inputs in the utility functions of the alternatives transit routes available to the passengers.

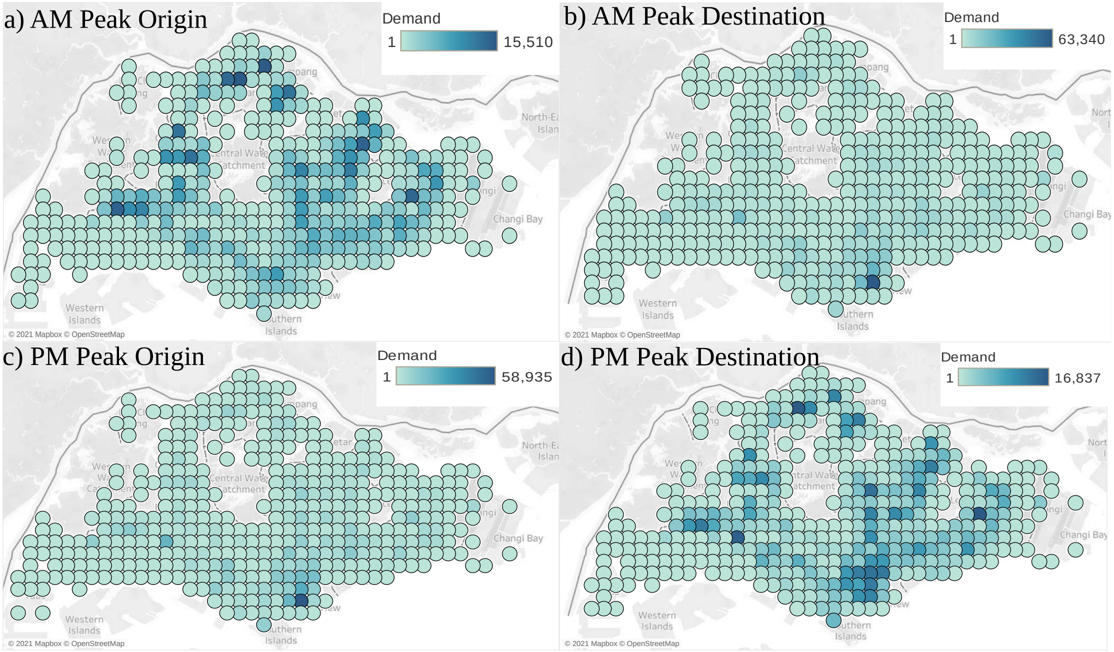


Figure 7: The spatial distribution of total travel demand for taxi and transit by a) AM peak origin locations b) AM peak destination locations c) PM peak origin locations, and d) PM peak destination locations for a single weekday in Singapore in Feb 2018.

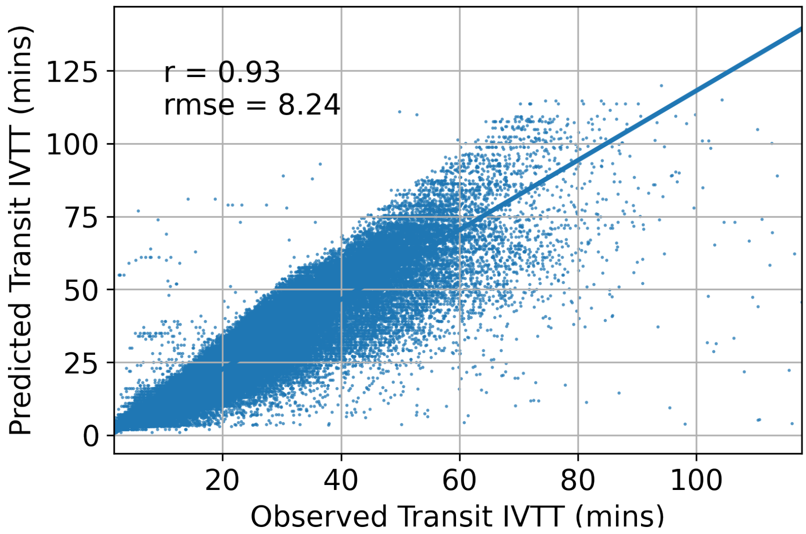
5.3. Taxi attribute generation results

The predictive performance of our taxi attribute generation procedure is given in Figure 9 with Pearson correlation coefficient, r and Root Mean Square Error, $rmse$ between the observed and predicted taxi attributes. It should be noted that since our algorithm determines the shortest path given the observed travel times for each time window, the route generated by our model can significantly differ from the actual travel path of the taxi driver on the ground. In addition, our network does not include traffic signals and traffic junctions; delays introduced by traffic signals and junctions are instead captured in the travel time matrix generated from the taxi GPS trajectory data set. The combination of the shortest path, and the different traffic conditions and network features along this path can result in predicted travel times that are often shorter than the observed travel times. Still our model has reasonably good performance in predicting travel time and travel distance for a taxi journey. Travel time values from the model is directly used in the taxi utility function to estimate the parameters of the choice model whereas the travel distance is used as an input to the taxi fare model.

Taxi fare model is a simple linear regression model and the model with travel distance as the exploratory variable provided a good fit for the data. We also explored a model with travel time as the exploratory variable, but the model with travel distance provided better fit and predictions and hence was selected. The selected model is shown in Equation 10.

$$Fare_{taxi} = \beta_c + \beta_{distance} \times (distance)\,, \tag{10}$$

where $Fare_{taxi}$ is the taxi fare in S\$, $distance$ is the taxi travel distance in kilometers, β_c and $\beta_{distance}$ are the estimated parameters. The Ordinary Least Square (OLS) estimation results for the parameters are shown in Table 1. The predictive performance of the taxi fare model is also reasonably good for our application and is shown in Figure 9. Taxi fares from this model is used in the taxi utility function for the choice model along with the taxi travel time.



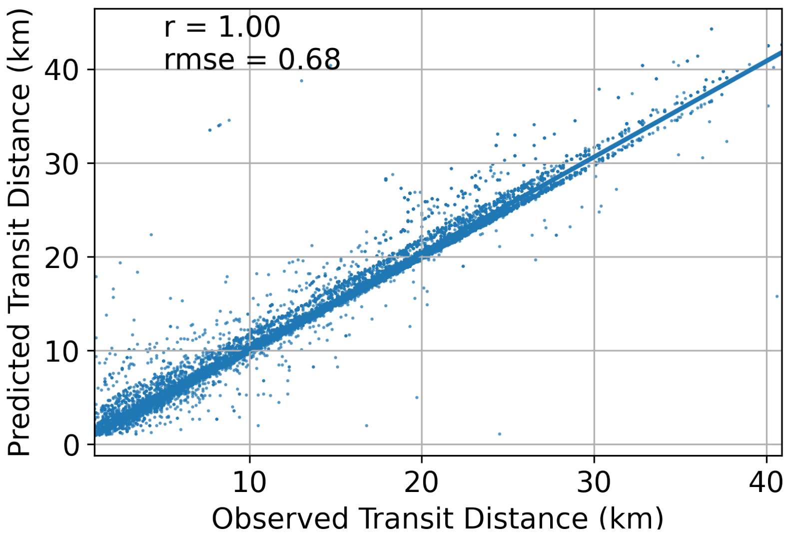


Figure 8: Goodness of fit plots between model predictions vs observed for a) transit in-vehicle travel time (IVTT) and b) transit distance.

Table 1: Parameter estimates of the regression model for taxi fare

Parameter description	Estimated values	Std. Error	t-stat
Intercept, β_c	5.375	0.012	436.072
distance, $\beta_{distance}$	0.826	0.001	898.587
Summary statistics			
Number of observations = 209615			
Number of estimated parameters = 2			
R-squared = 0.794			
Adjusted R-squared = 0.794			

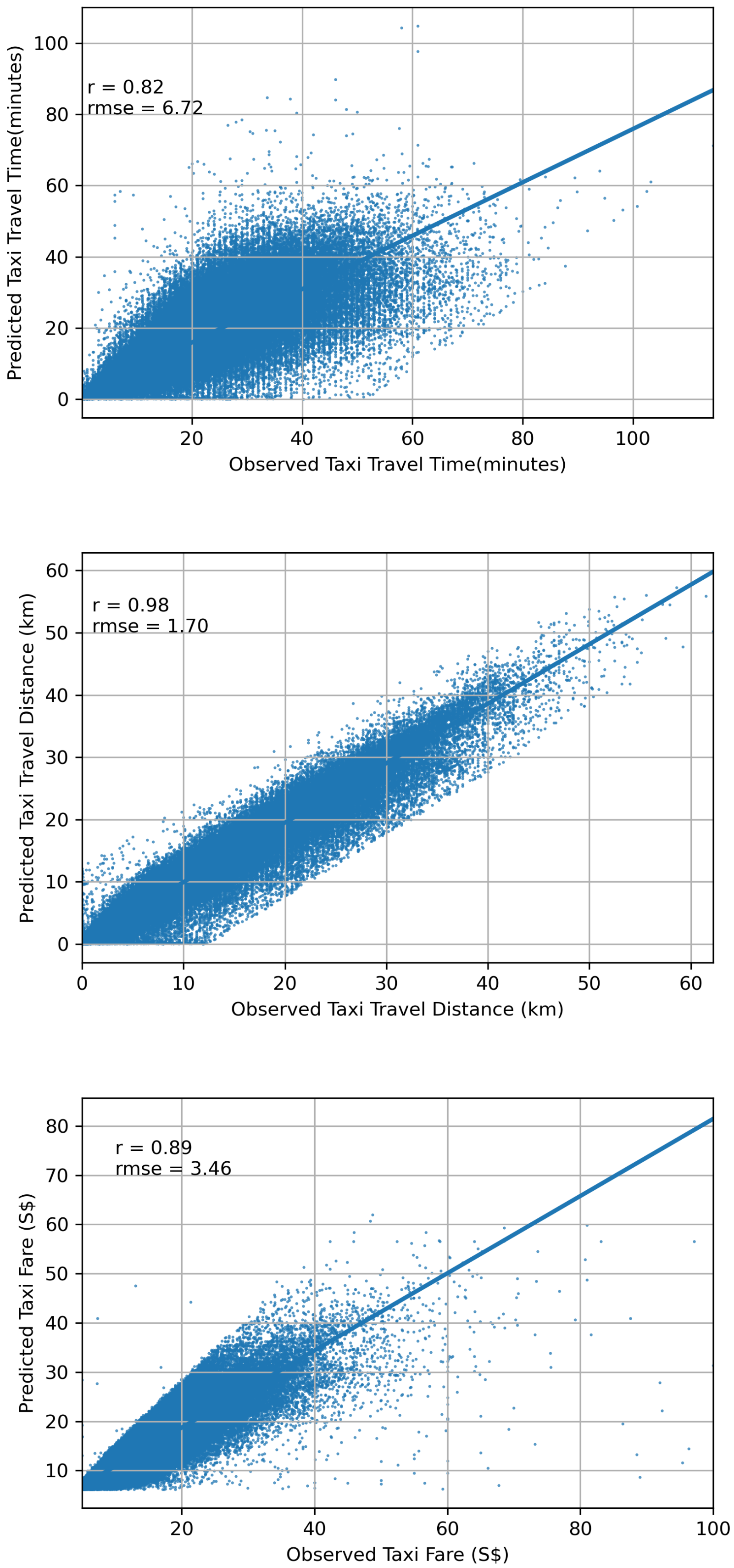


Figure 9: Performance of taxi attribute generation models

5.4. Choice model estimation results

The estimation results for the two-level nested logit and cross-nested logit models are is shown in Table 2. The data used in the model estimation was derived using weighted random sampling procedure from a single weekday data, where the weights for transit and taxi was based on the respective observed mode share for captive riders as explained in Section 4.1, making it a modally representative sample. We did not give spatial weights, but as a check for this, the model was re-estimated using different random samples and at different sample sizes. The sample sizes were gradually increased up to 1 million to see the variance in the parameter estimates. Although the results presented here are estimated from 1 million samples, we note that there is not much change in these estimates when sample sizes are varied from 1000 to 1 million. Consistent parameter estimates across different samples showed that the data used in our our model is spatially representative.

The estimated coefficients have the expected signs and their statistical significance is high with high t-values in both models. The rho-squared value for both models are high with cross-nested logit model having slightly higher rho-squared value than that of the two-level nested logit.

The absolute value for the coefficient corresponding to the number of transfers is higher compared to other variables in both models, which shows that passengers prefer routes with lesser number of transfers and they put highest importance on number of transfers compared to other attributes of the route while comparing the routes. Similar results are observed in Kim et al. (2020), Hensher and Rose (2007) and Raveau et al. (2014). Walk time for transfer is the variable with next highest value for coefficient, followed by fare and travel time in the two-level nested logit model. The cross-nested logit model with mode specific IVTT and Fare coefficient shows that there is higher negative utility for PT fare compared to taxi fare, which shows the lower willingness to pay for the transit alternatives compared to taxi.

The mode specific constant for the taxi option β_{Taxi} is estimated separately for peak hour (β_{Taxi}^{Peak}) and off-peak hours ($\beta_{Taxi}^{OffPeak}$). Peak and off-peak coefficients for taxi have negative values, indicating that the passengers generally prefer public transit over taxi. The taxi constant for peak has higher negative value than the taxi constant for off-peak, which shows that passengers have a higher chance of taking taxis during off-peak hours than peak hours. A major portion of the journeys during peak hours are likely to be work related journeys, whereas the majority of off-peak journeys are likely to be non-work related. Based on this assumption we can make one interpretation that the passengers have slightly higher preference for taxi for non-work related journeys compared to work related journeys. The scale parameter for taxi nest, μ_{Taxi} is set to 1 to allow the estimation of the rest of the nest scale parameters in both models. The $IVTT$ coefficient for both PT alternatives and Taxi are very similar in the cross-nested logit model suggesting that the influence of IVTT on the choice is similar regardless of the mode. But, the $Fare$ coefficient for PT has higher negative value compared to that of the taxi. This clearly shows that the travellers perceive the fare for PT and taxi differently and they are more sensitive to PT fare than taxi fare.

To further analyze the relative importance of the variables, we found the ratios of coefficient of each variable to the coefficient of fare, which are the willingness to pay (WTP) estimates. As cost coefficients for PT and taxi were found to be different in Model 2, the WTP estimates are derived using Model 2 and is presented in Table 3. The willingness to pay values show that passengers are willing to pay 0.05S$ for one minute reduction in in-vehicle travel time in PT and 0.32S$ for one minute reduction in in-vehicle travel time in taxi. This clearly shows the much lower value of time while travelling in PT compared to Taxi. For one minute reduction in walk time for transfers,

Table 2: Estimation results for the nested mode and route choice models

Parameter description	Estimated values	Std. Error	t-stat
Model 1: two-level nested logit model with generic utility parameters			
In-vehicle travel time (in mins), β_{IVTT}	-0.0556	0.000258	-216
Number of transfers, β_{NoT}	-2.750	0.0119	-230
Walk time for transfers (in mins), β_{WT}	-0.207	0.00236	-87
Fare (in S$), β_{Fare}	-0.120	0.000870	-138
Taxi peak, β_{Taxi}^{Peak}	-2.304	0.0115	-200
Taxi off peak, $\beta_{Taxi}^{OffPeak}$	-1.583	0.00927	-170
Scale parameter for PT nest, μ_{PT}	1.865	0.00819	228
Summary statistics			
Number of observations = 1000000			
Number of estimated parameters = 7			
Initial log-likelihood = -2640199			
Final log-likelihood = -961094			
Rho-squared = 0.636			
Rho-squared bar = 0.636			
Model 2: Cross-nested logit model with mode specific utility parameters			
In-vehicle travel time for PT (in mins), β_{IVTT}^{PT}	-0.0524	0.000308	-170
In-vehicle travel time for Taxi (in mins), β_{IVTT}^{Taxi}	-0.0522	0.00135	-38
Number of transfers, β_{NoT}	-3.0269	0.00799	-379
Walk time for transfers (in mins), β_{WT}	-0.343	0.00308	-111
Fare for PT (in S$), β_{Fare}^{PT}	-1.131	0.0212	-53
Fare for taxi (in S$), β_{Fare}^{Taxi}	-0.166	0.00222	-75
Taxi off peak, $\beta_{Taxi}^{OffPeak}$	-2.889	0.0198	-145.968
Taxi peak, β_{Taxi}^{Peak}	-3.634	0.0202	-180
Scale parameter for bus nest, μ_{Bus}	1.988	0.00679	293
Scale parameter for train nest, μ_{Train}	1.702	0.00980	173
Nest membership parameter for bus nest, α_{Bus}	0.404	0.000774	522
Summary statistics			
Number of observations = 1000000			
Number of estimated parameters = 11			
Initial log-likelihood = -2894279			
Final log-likelihood = -931681			
Rho-squared = 0.678			
Rho-squared bar = 0.678			

Table 3: Willingness to pay estimates

Variables	Current study	Hess et al. (2017)
PT IVTT (in mins)	0.05	0.11-0.24
Taxi IVTT (in mins)	0.32	0.55
Number of Transfers	2.68	0.41-0.68
Walk time (in mins)	0.30	0.16-0.23

Table 4: Predicted mode shares under selected scenarios

Scenario	PT fare	Taxi fare	PT share (%)	Taxi share (%)
B	No change	No change	89.6	10.4
ZPTF	0	No change	95.8	4.2
PTFI	2.2% increase	No change	89.4	10.6
TFI	No change	4%increase	90.1	9.9
PTTFI	2.2% increase	4%increase	89.4	10.6

passengers are willing to pay 0.30S$. As expected from the parameter value, highest willingness to pay was found to be for reducing the number of transfers which is 2.68S$. Table 3 also shows the WTP values from a large scale stated preference survey study conducted in Singapore (Hess et al. (2017)). A range of willingness to pay values for each variable discussed here was estimated for different modes in Hess et al. (2017) under different conditions of crowdedness. Value of IVTT for both PT and taxi modes have slightly lower values in this study compared to Hess et al. (2017), whereas the value of number of transfers and walk time was found to be higher in this study compared to Hess et al. (2017), with larger difference in values for the number of transfers variable. The difference in the WTP estimates may be due to the well-acknowledged differences in behaviors observed between stated and revealed preference data.

6. Application

The model presented in this study can be used in predicting the mode split between transit and taxi when service attributes of these alternatives vary. We demonstrate one such application by simulating selected scenarios with changes in transit and taxi fares. A test sample, which is different from the estimation sample was drawn for the policy testing. We considered the following scenarios - baseline scenario (B), Zero PT Fare (ZPTF) scenario , PT Fare Increase (PTFI) scenario, Taxi Fare Increase (TFI) scenario and PT and Taxi Fare Increase (PTTFI) scenario. The percentage changes in PT and taxi fare in the fare increase scenarios are based on the latest fare hikes for these modes in Singapore. Details of the scenarios and the predicted mode shares under each scenario are shown in Table 4.

The baseline mode shares predicted by the model was 89.6% for PT and 10.4% for taxi, whereas the observed mode shares for captive riders were found to be 11% for taxi and 89% for PT, showing that the model is able to predict the observed mode shares accurately in an out-of-sample prediction. The PT mode share under the zero PT fare scenario was found to be 95.8%. This shows that the maximum PT mode share that can be achieved by reducing PT fare is 95.8%, given other factors influencing the mode choice remains unchanged. The increase of PT fare by 2.2% resulted in a nominal reduction in PT mode share. Similarly 4% increase in taxi fare resulted in nominal increase in PT mode share. The scenario where PT fare was increased by 2.2% and taxi fare was increased by 4%, there is a slight decrease in the PT mode share. Such analysis are valuable for transport policy makers while making fare adjustments.

7. Conclusions

The mode and transit route choice behaviour of passengers on large-scale multi-modal public transport network is studied in this paper with nested and cross-nested logit models that are estimated using smart card and taxi data from Singapore. A novel data fusion approach is used in the study to combine the smart card and taxi data that are of different geographical resolutions. Large volume of revealed preference data on travel behaviour is available in transaction data sets like smart card data and taxi data and the approach used in this work demonstrates how such data sources can be combined and used for predicting travellers' choice of non-private modes and transit routes.

The utility function of the estimated nested logit model model has in-vehicle travel time, number of transfers, walk time for transfers, and fare as exploratory variables. The estimation results show that by including these variables the model is able to achieve good statistical fit thereby explaining the choice behaviour of passengers well. The results show that passengers perceive number of transfers with very high negative utility compared to other variables in the model. Similarly, walk time for transfers are perceived with higher negative utility than the in-vehicle travel time. Willingness to pay estimates were calculated and compared against other estimation results from Singapore.

Unlike other transit mode and route choice models found in the literature, our model considers all non-private modes of transport including taxi. An application of the model by changing the taxi and transit fares are shown in the study. Due to the dis-aggregate nature of the model, the estimated model from this study can be used in large scale agent-based simulators to model the mode and transit route choice behaviour of agents. Such an agent-based simulation model will help in testing the system level impacts of changes in the public transport network. Examples for changes in transit network include addition of a new MRT line or service disruptions in existing MRT lines. The models developed in this study is used in Othman et al. (2022) to represent the mode and route choice behaviour of passengers.

We also acknowledge some of the limitations of the work. In the current study, walking is assumed to be the access and egress mode for public transport and trip chaining of taxi with other public transport modes is not considered due to the unavailability of data. Similarly, the smart card and taxi data have limited or no information on the passenger's characteristics and hence are not considered in our model. Another limitation is the exclusion of private modes in the analysis. Although private modes are excluded, the model presented in the study can be used in predicting the behavior of passengers who do not have other alternatives in their choice set. For places like Singapore, where the private car ownership is low and the slow modes like walking and biking are used only for very short trips, majority of the population are left with two alternatives in their choice set–transit or taxi–and the model developed in this study is useful under such conditions. Also, when the data from the other modes are available this model can be extended by adding another level of choice on top of the current model structure, where the first level of choice would be between private and non-private modes of transport.

Acknowledgements

This research is supported by the National Research Foundation, Singapore, and the Land Transport Authority under its Urban Mobility Grand Challenge Programme (Award No UMGC-L005). The views expressed herein are those of the authors and are not necessarily those of the funding agencies.

Author contributions

The authors confirm contribution to the paper as follows. **Rakhi Manohar Mepparambath:** Conceptualization, Methodology, Data curation, Formal analysis, Visualization, and Manuscript preparation. **Yong Sheng Soh**: Methodology, Data curation, Formal analysis, and Manuscript preparation. **Vasundhara Jayaramana**: Methodology, Data curation, Formal analysis, Visualization, and Manuscript preparation. **Hong En Tan**: Conceptualization, Methodology, Data curation, and Manuscript preparation. **Muhamad Azfar Ramli**: Conceptualization, Methodology, Manuscript preparation, Funding acquisition, and Supervision.

References

Alsger, A. A., Mesbah, M., Ferreira, L., and Safi, H. (2015). Use of smart card fare data to estimate public transport origin–destination matrix. *Transportation Research Record*, 2535(1):88–96. doi: `https://doi.org/10.3141/2535-10`.

Anderson, M. K., Nielsen, O. A., and Prato, C. G. (2017). Multimodal route choice models of public transport passengers in the greater copenhagen area. *EURO Journal on Transportation and Logistics*, 6(3):221–245. doi: `https://doi.org/10.1007/s13676-014-0063-3`.

Bierlaire, M. (2018). Pandasbiogeme: a short introduction. *EPFL (Transport and Mobility Laboratory, ENAC)*.

Bierlaire, M., Bolduc, D., and McFadden, D. (2008). The estimation of generalized extreme value models from choice-based samples. *Transportation Research Part B: Methodological*, 42(4):381–394. doi: `https://doi.org/10.1016/j.trb.2007.09.003`.

Cheng, Z., Trépanier, M., and Sun, L. (2020). Probabilistic model for destination inference and travel pattern mining from smart card data. *Transportation*, pages 1–19. doi: `https://doi.org/10.1007/s11116-020-10120-0`.

Eluru, N., Chakour, V., and 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(2):129–149. doi: `https://doi.org/10.1007/s12469-012-0056-2`.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Foell, S., Phithakkitnukoon, S., Kortuem, G., Veloso, M., and Bento, C. (2014). Catch me if you can: Predicting mobility patterns of public transport users. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1995–2002. IEEE. doi: `https://doi.org/10.1109/ITSC.2014.6957997`.

Geisberger, R., Sanders, P., Schultes, D., and Vetter, C. (2012). Exact routing in large road networks using contraction hierarchies. *Transportation Science*, 46(3):388–404. doi: `https://doi.org/10.1287/trsc.1110.0401`.

Goulet-Langlois, G., Koutsopoulos, H. N., Zhao, Z., and Zhao, J. (2017). Measuring regularity of individual travel patterns. *IEEE Transactions on Intelligent Transportation Systems*, 19(5):1583–1592. doi: `https://doi.org/10.1109/TITS.2017.2728704`.

Hensher, D. A. and Rose, J. M. (2007). Development of commuter and non-commuter mode choice models for the assessment of new public transport infrastructure projects: a case study. *Transportation Research Part A: Policy and Practice*, 41(5):428–443. doi: https://doi.org/10.1016/j.tra.2006.09.006.

Hess, S., Murphy, P., Le, H., and Leong, W. Y. (2017). Estimation of new monetary valuations of travel time, quality of travel, and safety for singapore. *Transportation Research Record*, 2664(1):79–90.

Hunt, J. D. (1990). A logit model of public transport route choice. *ITE Journal*, 60(12).

Hussain, E., Bhaskar, A., and Chung, E. (2021). Transit od matrix estimation using smartcard data: Recent developments and future research challenges. *Transportation Research Part C: Emerging Technologies*, 125:103044. doi: https://doi.org/10.1016/j.trc.2021.103044.

Jánošíková, L., Slavík, J., and Koháni, M. (2014). Estimation of a route choice model for urban public transport using smart card data. *Transportation planning and technology*, 37(7):638–648. doi: https://doi.org/10.1080/03081060.2014.935570.

Kang, M. J., Ataeian, S., and Amiripour, S. M. (2021). A procedure for public transit od matrix generation using smart card transaction data. *Public Transport*, 13(1):81–100. doi: https://doi.org/10.1007/s12469-020-00257-7.

Kieu, L.-M., Bhaskar, A., and Chung, E. (2015). A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card afc data. *Transportation Research Part C: Emerging Technologies*, 58:193–207. doi: https://doi.org/10.1016/j.trc.2015.03.033.

Kim, I., Kim, H.-C., Seo, D.-J., and Kim, J. I. (2020). Calibration of a transit route choice model using revealed population data of smartcard in a multimodal transit network. *Transportation*, 47(5):2179–2202. doi: https://doi.org/10.1007/s11116-019-10008-8.

Kumar, P., Khani, A., and He, Q. (2018). A robust method for estimating transit passenger trajectories using automated data. *Transportation Research Part C: Emerging Technologies*, 95:731–747. doi: https://doi.org/10.1016/j.trc.2018.08.006.

Kurauchi, F., Schmöcker, J.-D., Fonzone, A., Hemdan, S. M. H., Shimamoto, H., and Bell, M. G. (2012). Estimating weights of times and transfers for hyperpath travelers. *Transportation research record*, 2284(1):89–99. doi: https://doi.org/10.3141/2284-11.

Lam, S.-H. and Xie, F. (2002). Transit path-choice models that use revealed preference and stated preference data. *Transportation Research Record*, 1799(1):58–65. doi: https://doi.org/10.3141/1799-08.

Lathia, N., Smith, C., Froehlich, J., and Capra, L. (2013). Individuals among commuters: Building personalised transport information services from fare collection systems. *Pervasive and Mobile Computing*, 9(5):643–664. doi: https://doi.org/10.1016/j.pmcj.2012.10.007.

Lei, D., Chen, X., Cheng, L., Zhang, L., Ukkusuri, S. V., and Witlox, F. (2020). Inferring temporal motifs for travel pattern analysis using large scale smart card data. *Transportation Research Part C: Emerging Technologies*, 120:102810. doi: https://doi.org/10.1016/j.trc.2020.102810.

Li, M., Dong, L., Shen, Z., Lang, W., and Ye, X. (2017). Examining the interaction of taxi and subway ridership for sustainable urbanization. *Sustainability*, 9(2):242. doi: https://doi.org/10.3141/1645-17.

Liu, S., Yamamoto, T., Yao, E., and Nakamura, T. (2020). Exploring travel pattern variability of public transport users through smart card data: Role of gender and age. *IEEE Transactions on Intelligent Transportation Systems*. doi: https://doi.org/10.1109/TITS.2020.3043021.

Liu, X., Gong, L., Gong, Y., and Liu, Y. (2015). Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography*, 43:78–90. doi: https://doi.org/10.1016/j.jtrangeo.2015.01.016.

Liu, Y., Kang, C., Gao, S., Xiao, Y., and Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of geographical systems*, 14(4):463–483. doi: https://doi.org/10.1007/s10109-012-0166-z.

Lo, H. K., Yip, C.-W., and Wan, Q. K. (2004). Modeling competitive multi-modal transit services: a nested logit approach. *Transportation Research Part C: Emerging Technologies*, 12(3-4):251–272. doi: https://doi.org/10.1016/j.trc.2004.07.011.

LTA (2021). Land Transport DataMall. Available at https://datamall.lta.gov.sg/content/datamall/en/dynamic-data.html, Accessed 20.09.2021.

Ma, X., Liu, C., Wen, H., Wang, Y., and Wu, Y.-J. (2017). Understanding commuting patterns using transit smart card data. *Journal of Transport Geography*, 58:135–145. doi: https://doi.org/10.1016/j.jtrangeo.2016.12.001.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., and Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12. doi: https://doi.org/10.1016/j.trc.2013.07.010.

Munizaga, M. A. and Palma, C. (2012). Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C: Emerging Technologies*, 24:9–18. doi: https://doi.org/10.1016/j.trc.2012.01.007.

Nassir, N., Hickman, M., and Ma, Z.-L. (2019). A strategy-based recursive path choice model for public transit smart card data. *Transportation Research Part B: Methodological*, 126:528–548. doi: https://doi.org/10.1016/j.trb.2018.01.002.

Othman, N. B., Jayaraman, V., Chan, W., Loh, K., Rajendram, R., Mepparambath, R. M., Agrawal, P., Ramli, M. A., and Zheng, Q. (2022). Summit: A multi-modal agent-based co-simulation of urban public transport with applications in contingency planning. Available at http://dx.doi.org/10.2139/ssrn.4258155.

PTC (2019). Bus and train fare structure. Available at https://www.ptc.gov.sg/fare-regulation/bus-rail/fare-structure, Accessed 26.09.2021.

Pursula, M. and Weurlander, M. (1999). Modeling level-of-service factors in public transportation route choice. *Transportation Research Record*, 1669(1):30–37. doi: https://doi.org/10.3141/1669-04.

Raveau, S., Guo, Z., Muñoz, J. C., and Wilson, N. H. (2014). A behavioural comparison of route choice on metro networks: Time, transfers, crowding, topology and socio-demographics. *Transportation Research Part A: Policy and Practice*, 66:185–195. doi: https://doi.org/10.1016/j.tra.2014.05.010.

Raveau, S., Muñoz, J. C., and De Grange, L. (2011). A topological route choice model for metro. *Transportation Research Part A: Policy and Practice*, 45(2):138–147. doi: https://doi.org/10.1016/j.tra.2010.12.004.

Tang, J., Liu, F., Wang, Y., and Wang, H. (2015). Uncovering urban human mobility from large scale taxi gps data. *Physica A: Statistical Mechanics and its Applications*, 438:140–153. doi: https://doi.org/10.1016/j.physa.2015.06.032.

URA (2014). Master plan . Available at https://www.ura.gov.sg/Corporate/Planning/Master-Plan, Accessed 30.09.2021.

Van Der Hurk, E., Kroon, L., Maróti, G., and Vervest, P. (2014). Deduction of passengers' route choices from smart card data. *IEEE Transactions on intelligent transportation systems*, 16(1):430–440. doi: https://doi.org/10.1109/TITS.2014.2333583.

Vovsha, P. and Bekhor, S. (1998). Link-nested logit model of route choice: overcoming route overlapping problem. *Transportation research record*, 1645(1):133–142. doi: https://doi.org/10.3141/1645-17.

Yap, M., Cats, O., and van Arem, B. (2020). Crowding valuation in urban tram and bus transportation based on smart card data. *Transportmetrica A: Transport Science*, 16(1):23–42. doi: https://doi.org/10.1080/23249935.2018.1537319.

Zhan, X., Hasan, S., Ukkusuri, S. V., and Kamga, C. (2013). Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C: Emerging Technologies*, 33:37–49. doi: https://doi.org/10.1016/j.trc.2013.04.001.

Zhao, J., Zhang, F., Tu, L., Xu, C., Shen, D., Tian, C., Li, X.-Y., and Li, Z. (2016). Estimation of passenger route choice pattern using smart card data for complex metro systems. *IEEE Transactions on Intelligent Transportation Systems*, 18(4):790–801. doi: https://doi.org/10.1109/TITS.2016.2587864.

Zhao, Z., Koutsopoulos, H. N., and Zhao, J. (2018). Individual mobility prediction using transit smart card data. *Transportation research part C: emerging technologies*, 89:19–34. doi: https://doi.org/10.1016/j.trc.2018.01.022.

Zhong, C., Manley, E., Arisona, S. M., Batty, M., and Schmitt, G. (2015). Measuring variability of mobility patterns from multiday smart-card data. *Journal of Computational Science*, 9:125–130. doi: https://doi.org/10.1016/j.jocs.2015.04.021.

Zhu, Y., Ozbay, K., Xie, K., and Yang, H. (2016). Using big data to study resilience of taxi and subway trips for hurricanes sandy and irene. *Transportation research record*, 2599(1):70–80. doi: https://doi.org/10.3141/2599-09.

© <2023>. This manuscript version is made available under the CC-BY-NC-ND 4.0 license
https://creativecommons.org/licenses/by-nc-nd/4.0/