Using metro smart card data to model location choice of after-work activities
An application to Shanghai

Wang, Yihong; Correia, Gonçalo Homem de Almeida; de Romph, Erik; Timmermans, H. J.P.(Harry)

DOI
10.1016/j.jtrangeo.2017.06.010

Publication date
2017

Document Version
Accepted author manuscript

Published in
Journal of Transport Geography

Citation (APA)
Wang, Y., Correia, G. H. D. A., de Romph, E., & Timmermans, H. J. P. (2017). Using metro smart card data to model location choice of after-work activities: An application to Shanghai. *Journal of Transport Geography*, *63*, 40-47. https://doi.org/10.1016/j.jtrangeo.2017.06.010

Using metro smart card data to model location choice of after-work activities: An application to Shanghai

Yihong Wang[a,*], Gonçalo Homem de Almeida Correia[a], Erik de Romph[a,b], H.J.P. (Harry) Timmermans[c]

[a] *Department of Transport and Planning, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands*

[b] *TNO, Delft, The Netherlands*

[c] *Department of the Built Environment, Section of Urban Systems & Real Estate, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands*

Abstract

A location choice model explains how travellers choose their trip destinations especially for those activities which are flexible in space and time. The model is usually estimated using travel survey data; however, little is known about how to use smart card data (SCD) for this purpose in a public transport network. Our study extracted trip information from SCD to model location choice of after-work activities. We newly defined the metrics of travel impedance in this case. Moreover, since socio-demographic information is missing in such anonymous data, we used observable proxy indicators, including commuting distance and the characteristics of one's home and workplace stations, to capture some interpersonal heterogeneity. Such heterogeneity is expected to distinguish the population and better explain the difference of their location choice behaviour. The approach was applied to metro travellers in the city of Shanghai, China. As a result, the model performs well in explaining the choices. Our new metrics of travel impedance to access an after-work activity result in a better model fit than the existing metrics and add additional interpretability to the results. Moreover, the proxy variables distinguishing the population seem to influence the choice behaviour and thus improve the model performance.

Keywords: Public transport; smart card data; location choice modelling; discrete choice model; demand forecast; transport planning.

1. Introduction

Travel behaviour is becoming more diverse and complex especially in large metropolitan areas. One of the most significant changes is that non-commuting travel demand takes a larger share than ever before (e.g., Lu and Gu, 2011). Therefore, the task of observing and analysing non-commuting travel demand is becoming important today. This task is not only relevant for transport planners to better understand movements of travellers, but also for service and retail business planners to understand where people would like to consume and where their customers come from (Sivakumar and Bhat, 2007). Moreover, economists regard the accessibility to non-commuting activities as an important indicator to reflect

* Corresponding author. Tel: +31 (0)61 68 45160.
E-mail addresses: Y.Wang-14@tudelft.nl (Y. Wang), G.Correia@tudelft.nl (G. Correia), E.deRomph@tudelft.nl (E. de Romph), h.j.p.timmermans@tue.nl (H.J.P. Timmermans).

quality of life (Nakamura et al., 2016; Suriñach et al., 2000). These relevant perspectives have led the transportation research field to expand its scope to topics like accessibility (Dong et al., 2006), social exclusion (Schönfelder and Axhausen, 2003), subjective well-being (De Vos et al., 2013), etc., in addition to traditional transport problems particularly focusing on network levels of service.

To cope with the increasing non-commuting demand, the usage of public transport (PT) to access retail and service facilities has been encouraged in many cities due to the concentration of people (Castillo-Manzano and López-Valpuesta, 2009; Ibrahim and McGoldrick, 2003). Urban decision makers need to know where large recreational centres should be located and how PT network should be planned to meet the considered objectives. Answering these questions requires the prediction of non-commuting OD matrices in many "what-if" scenarios, based on the understanding of people's activity-travel behaviour including, but not limited to, location choice. A relevant and interesting perspective is the activity-based travel demand modelling, which focuses on individuals and regards travelling as the result of the need to participate in activities (Rasouli and Timmermans, 2014). However, few studies have adopted this methodology focused on PT network. In this paper, we aim to fill this gap by using new available travel demand data sources, namely, smart card data (SCD). We focus on travel demand of after-work activities since it is a significant part of non-commuting travel demand especially on weekdays (Demerouti et al., 2009). Our research can also be regarded as a complement to the existing research that uses SCD to study commuting patterns (Ma et al., 2017; Zhou et al., 2014).

Compared to traditional mobility survey data, SCD have several advantages and disadvantages to reveal how people travel by PT (Bagchi and White, 2005; Pelletier et al., 2011). Firstly, collecting such data is more efficient, saving both time and money, compared to large-scale surveys. Secondly, SCD usually correspond to a larger sample and the observations can be longitudinal in time (Morency et al., 2007). On the other hand, trip purpose is difficult to obtain in SCD and needs to be estimated using other methods (Devillaine et al., 2013; Kuhlman, 2015; Long et al., 2012). In some cases, destination information needs to be estimated as well because some PT networks do not request a check-out (Trépanier et al., 2007). The very relevant personal socio-demographic information is most of the times not available for confidentiality reasons which decreases the possibility to do a more thorough analysis of particular behavioural traits of the population (Pelletier et al., 2011).

The advantages of using SCD have allowed researchers to obtain more accurate estimates of transit demand, which have led to many applications. Using the data collected during 277 consecutive days, Morency et al. (2007) examined the variability of transit use. Some studies proposed to cluster and classify the regularity of transit travel patterns by mining SCD (Goulet Langlois et al., 2016; Ma et al., 2013). Estimating origin-destination (OD) transit trip matrices is a usual application of SCD (Munizaga and Palma, 2012). It can further serve as a fixed input to passenger flow assignment (Sun et al., 2015), OD flow visualization (Liu et al., 2009; Long et al., 2012) and any other post hoc analysis, such as commuting efficiency assessment (Zhou et al., 2014). However, only a few attempts have been made to use SCD to build explanatory trip distribution or location choice models, in order to predict the OD

matrices as a result of the changes made to transport systems and land use. One example is the gravity model developed by Goh et al. (2012) to understand aggregate commuting OD flows by metro. We believe that not only the characteristics of SCD but also the research objective in our study is a better fit for a disaggregate activity-based travel demand modelling framework.

In this study, we use SCD to model location choice of after-work activities. The innovation of our approach firstly lies in the creation of new metrics to model travel impedance in location choice of after-work activities. Secondly, this is the first time that proxy variables, which can be observed in anonymous SCD, are used to capture some interpersonal heterogeneity in order to explain the difference of their location choice behaviour. Thanks to the Shanghai Open Data Apps (SODA) contest[1], a full-population dataset of one-month PT smart card transaction records for the city of Shanghai (China) was made available, allowing us to explore this methodology in a large-size real-world case scenario.

This paper is organized as follows. First, the methodology is described. Then, the data of Shanghai is further explained. Following that, we present the application of our method. In the final section, we take conclusions and point out directions for future research.

2. Methodology

We start by defining the scope to which our methodology can be applied. The method can be applied in a metro network composed of stations with services connecting them, where the automated fare collection system forces travellers to check in and check out at the stations where they board and alight respectively. Therefore, the following information of each trip is available through SCD: anonymous identity (ID) of the user, IDs of boarding and alighting stations and timestamp. A trip is defined to start from an origin station near which the previous activity has been finished, and end at a destination station where the next activity will take place. In our case, the recorded boarding and alighting stations are not necessarily an origin or a destination station of a trip. In other words, a trip including any transfers should not be regarded as two separate ones. Moreover, a daily trip chain is the ordered set of trips done by an individual within one day.

2.1 Detecting commuters

Several studies have been performed on the detection of commuters as well as their home and workplace stations from SCD (Chakirov and Erath, 2012; Long and Thill, 2015). By recurring to travel survey data, researchers have either predefined the rules or trained the models to predict if a smart card user is a commuter and if the purpose of a PT trip recorded in SCD is home, work or other, based on several observed factors, such as activity start time. In our method, we used a similar principle for activity identification, but due to the unavailability of travel survey data, we predefined the rules with the parameters identified in the literature.

We used the following rule applied by Long et al. (2012) to determine one's home station: any boarding station of the first trip done by an individual on a weekday was defined as a so-

[1] http://soda.datashanghai.gov.cn/ (retrieved date: November 21st, 2015)

called candidate home station of this individual, and the station appearing most frequently as a candidate home station during the observed period was defined as the definitive home station of this individual. There could be more than one station appearing most frequently. In such cases, Long et al. (2012) compared the land use around the stations and assigned the station in a more residential environment to be the definitive home station.

In SCD, activity duration can approximately be regarded as the time gap between a check-out and the subsequent check-in at the same station when the access and egress mode is walking. If the activity duration of visiting a station was longer than 6 hours on a weekday, we identified the station as a so-called candidate workplace station. Long et al. (2012) selected this parameter based on the travel survey data from Beijing, China, and thus we think that it is the best reference for our study of Shanghai despite the differences between the two cities. Next, the station appearing most frequently as a candidate workplace station during the observed period was defined as the definitive workplace station. If there were more than one station appearing most frequently, we calculated for each station the distance from home multiplied by the frequency of visits during the observed period, as suggested by Alexander et al. (2015), and the station with the largest product was defined as the definitive workplace station.

Commuters were defined as those who had both detected definitive home and workplace stations. Due to access and egress, home and workplace stations are not, in many cases, the real locations of home and workplace but can be regarded as proxies for those, especially when the access and egress mode is walking.

One drawback of our method is that those commuters who have multiple home or workplace stations or have flexible working hours are difficult to detect. If necessary and possible, we recommend a more flexible approach relying on travel survey data. However, this step is not the main focus of our work, and our current method using the parameters identified in the literature is sufficient to detect a great number of commuters whom we can study regarding their after-work station choice behaviour.

2.2 Extracting individual daily metro trip chains

We assume that within one day, travellers do an activity between every two consecutive trips, and the purpose of this activity can be estimated based on the check-out station of the former trip and the check-in station of the latter. If they are the same one, the purpose can be classified into home, work or secondary activity dependent on whether the station is the home station, the workplace station or neither for that individual; if they are different due to the interim unobservable movement by using other modes, we do not classify any activity purpose. Note that the first activity on one day is dependent only on the check-in station of the first trip, and the last activity is dependent only on the check-out station of the last trip.

The diagram of an individual daily metro trip chain starts in the first activity within a day, represented as a node, connected by an edge representing the trip to the second activity, connected sequentially until the last activity. An example is shown in Figure 1, where each activity is labelled with its type and the grey box indicates where the chain starts. The commuter first travels from the home station to the workplace station at 8:00 and stays at

the workplace station until 17:30. After staying at another station for 90 minutes, this person checks out there and travels back home.

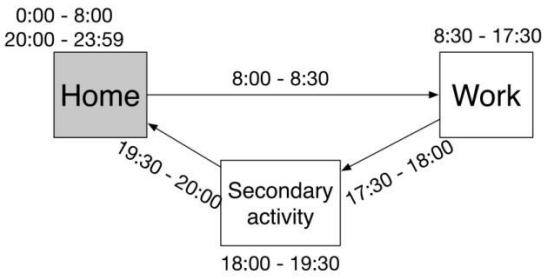


Figure 1 An example of an individual daily PT trip chain

2.3 Modelling station choices for after-work activities

In this paper, we focus on modelling station choice of metro commuters for after-work activities. Location choice involves a trade-off between attractiveness and travel impedance. We assumed that the attractiveness of a station for after-work activities is time-invariant. Travel impedance is a function of PT travel time, PT network distance, PT costs and number of PT transfers. In existing location choice models, there were three ways to model travel impedance to perform a secondary activity in a trip chain. The traditional way was to consider only the impedance of travelling between the activity location and home (Arentze and Timmermans, 2004). However, Arentze and Timmermans (2007) found that this measurement would result in the overestimation of the impedance between locations of activities within trip chains, and they proposed the concept of detour travel impedance:

$$DT_s = d(O_s, s) + d(s, D_s) - d(O_s, D_s) \quad (1)$$

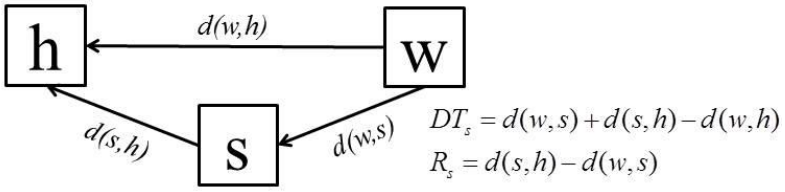
In this equation, O_s is the origin of the trip to a candidate location s for the secondary activity, and D_s is the destination of the trip from s. $d(x, y)$ is the travel impedance from x to y.

Despite the wide use of this concept in existing travel demand models, such as MATSim (Horni, 2013), a disadvantage of this method is that it is not very sensitive in differentiating between distance from workplace or to home. Thus, while the previous definitions were adequate in the specific contexts of those studies, for our problem, it may be better to account for the effect of proximity to workplace vs. home. We defined the new metrics by complementing the detour impedance DT_s with a new variable R_s:

$$R_s = d(s, D_s) - d(O_s, s) \quad (2)$$

Table 1 summarizes the three possible ways to model travel impedances to perform an after-work activity in a trip chain. h, w and s represent home station, workplace station and candidate station for an after-work activity respectively, and the former two are respectively equivalent to the succeeding activity location D_s and the preceding activity location O_s in our specific case.

Table 1 Three ways to consider travel impedance in the choice of a location for an after-work activity

	Existing metrics		New metrics
Measurement	Home-based impedance $d(s,h)$	Detour impedance DT_s	Detour impedance DT_s and proximity to workplace vs. home R_s
Reference	Arentze and Timmermans (2004)	Arentze and Timmermans (2007); Horni (2013)	The approach in this study
Diagram			

Although we focus on a metro network, attention should be paid to other modes like the access and egress to trips made in the metro network. In this study, we only model the trips to perform after-work activities with walking as access and egress, and we assume that the generalized travel cost of walking access and egress is minor compared to the main part of the metro trip.

The characteristics of activities (i.e., activity start time and activity duration) can be inserted in the model to describe contexts of choice occasions. The underlying assumption, in line with existing travel demand models (Balmer et al., 2008), is that people have already generated their activity schedules before making location choices. Attributes related to individuals are generally missing in SCD; however, in our study, we proposed to use commuting distance and characteristics of home and workplace stations as proxies for the attributes of the travellers. Aggregating the number of people living and working near each station can help identify whether a station is categorized into a mainly residential area or a mainly commercial area (Liu et al., 2009). This can serve as a way to characterize each traveller's home and workplace stations.

Considering that choice making may also rely on the previously made choices, we include the effect of last choice feedback (i.e., first-order state dependence) in our model. Following the approach of Danalet et al. (2016), we estimate the model where the previous choice can be assumed to be strictly exogenous to the estimation. Danalet et al. (2016) also addressed a more advanced approach to deal with the initial conditions problem and related endogeneity bias in estimation. However, the consideration of these issues is beyond the scope of our paper. For the same reason, we do not consider time-variant attributes of alternatives and unobserved inter-individual and intra-individual response heterogeneity.

We used a discrete choice model to explain the station choice for after-work activities with the referred impedance structures in our study. Consider that an individual user u in the network of the study area is associated with the home station h_u and the workplace station w_u, where $h_u, w_u \in \boldsymbol{N}$, and $\boldsymbol{N}$ is the set of metro stations in an area. In addition, u is observed to have a set of choice occasions $\boldsymbol{J_u}$ over time. The choice set of the destinations for after-work activities is denoted as $\boldsymbol{S_{uj}} = \boldsymbol{R_{uj}} \setminus \{h_u, w_u\}$, where $\boldsymbol{R_{uj}}$ is the reachable subset

of $\boldsymbol{N}$ for u on choice occasion j. $\boldsymbol{R_{uj}}$ was calculated based on the following space-time constraints: (1) a commuter should not leave work earlier than the work schedule allows; (2) a commuter should not miss the last metro back home; (3) given the previous constraints, travel times to reach an after-work activity should not affect the activity start time and the activity duration. For each individual, we calculated the earliest time of departure from work during the observed period as the threshold to apply the first constraint. The timetables of the metro line were used to apply the second constraint. Travel time between every two stations can be calculated by averaging over the trips according to the SCD.

The deterministic part of the utility function for an alternative s ($s \in \boldsymbol{S_{uj}}$) on choice occasion j ($j \in \boldsymbol{J_u}$) of decision maker u in one month is the following:

$$\begin{aligned} V_{usj} = {} & Z_s[\alpha + \sum_m (\delta_m X_{um}) + \sum_n (\phi_n C_{ujn})] \\ & + \sum_k \{T_{usk}[\beta_k + \sum_m (\omega_{km} X_{um}) + \sum_n (\eta_{kn} C_{ujn})]\} + \gamma SAME_{usj} \end{aligned} \quad (3)$$

The descriptions of all variables and parameters are presented in

Table 2. $\alpha + \sum_m (\delta_m X_{um}) + \sum_n (\phi_n C_{ujn})$ is a function representing the preference for station attractiveness Z_s, and $\beta_k + \sum_m (\omega_{km} X_{um}) + \sum_n (\eta_{kn} C_{ujn})$ is a function representing the preference for reducing travel impedance T_{usk}. Both functions incorporate the effects of user-specific attributes X_{um} and activity characteristics C_{ujn} on taste variation. Therefore, the preferences vary across individuals and choice occasions (Sivakumar and Bhat, 2007). The specific indicators of T_{usk}, X_{um} and C_{ujn} are summarized in Table 3. The possible values of $SAME_{usj}$ under different conditions are given in the following equation:

$$SAME_{usj} = \begin{cases} 1 & \text{if station } s \text{ is chosen by individual } u \text{ on choice occation } j-1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Table 2 Variables and parameters in the deterministic utility function

Parameters		Variables	
γ	Preference for maintaining the previous choice	$SAME_{usj}$	Variable indicating the previous choice feedback
α	Baseline preference for attractiveness of station s	Z_s	Attractiveness of station s
β_k	Baseline preference for reducing the type k travel impedance	T_{usk}	The type k travel impedance associated with home and workplace stations of individual u and station s
δ_m	The extent of the preference for attractiveness of station s that can be captured by the attribute m of travelers	X_{um}	Variable for the attribute m of individual u
ϕ_n	The extent of the preference for attractiveness of station s that can be captured by the characteristic n of activities	C_{ujn}	Variable for the characteristic n of the activity performed by individual u on choice occasion j
ω_{km}	The extent of the preference for reducing the type k travel impedance that can be captured by the attribute m of travelers		
η_{kn}	The extent of the preference for reducing the type k travel impedance that can be captured by the characteristic n of activities		

Table 3 Indicators of travel impedance, user-specific attributes and activity characteristics in the utility function

Variables		Specific indicators
Travel impedance variables	Home-based impedance	$T_{us1} = d(s,h)$
	Detour impedance	$T_{us1} = DT_s$
	Detour impedance and home vs. workplace proximity	$T_{us1} = DT_s$ $T_{us2} = R_s$
User-specific attributes		X_{u1} : commuting distance X_{u2} : characteristics of home station X_{u3} : characteristics of workplace station
Activity characteristics		C_{uj1} : activity duration C_{uj2} : activity start time

Regarding the random part of the utility function, we used the spatially correlated logit model proposed by Bhat and Guo (2004) to consider the effect of spatial correlation between adjacent stations on the metro network. This is a cross-nested logit model (Train, 2009) with two characteristics: (1) it is a paired combinatorial logit model (Koppelman and Wen, 2000), and each paired nest includes a station and one of its adjacent station; (2) it defines the allocation parameters that reflect the degree to which each alternative belongs

to each nest. The probability of choosing an alternative can be calculated in a closed-form expression, where the dissimilarity parameter ρ ($0 < \rho \leq 1$) is designed to be equal across all paired nests and capture the general correlation between adjacent stations. There is no correlation between adjacent pairs of stations when $\rho = 1$, and the correlation increases as ρ decreases. In addition to the parameters in the deterministic part of the utility function, we need to estimate ρ as well. More details about the spatially correlated logit model can be found in Bhat and Guo (2004).

3. Background information and data of the case study

3.1 Study area

Shanghai is one of the most populated and fastest growing cities worldwide. The socio-economic development has influenced people's travel behaviour. Local travel surveys show that the trip generation rate of residents has increased in recent years. Meanwhile, the government invested in PT systems to mitigate traffic congestion led by the increasing private car ownership, resulting in an upward trend in the share of PT use (Lu and Gu, 2011). Among all PT modes, the Shanghai metro network is expanding the most in the last years. As shown in Figure 2, the metro system operates 14 metro lines, connecting 288 metro stations distributed in the region, among which there are 54 transfer stations (i.e., the stations where passengers can change from one line to another).

A shortest path algorithm can be used to calculate the shortest network distance between every two stations and the number of transfers along each of those paths. The trip fare is set by the operator based on the shortest network distance, and thus they are almost perfectly correlated. The perfect correlation also exists between travel time and network distance, since we assume that the speeds of metro service do not vary between different OD pairs. These are the reasons why in this application we did not use fare and travel time as components of generalized travel costs.

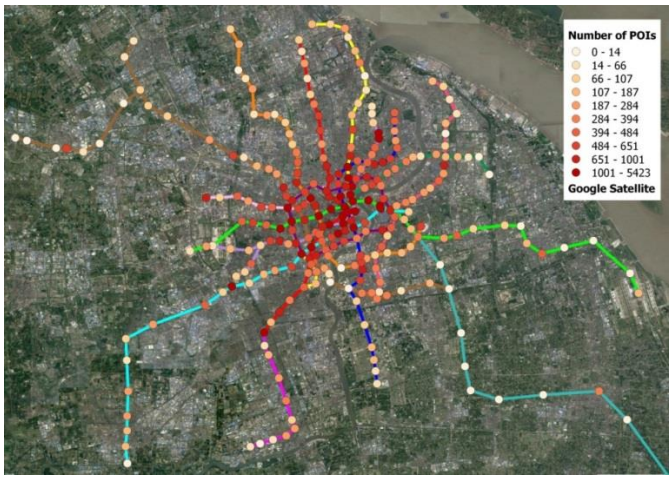


Figure 2 The metro network in Shanghai and number of POIs per station

On the website of Dianping[2], which is one of the most popular Chinese location-review services, we mined information of points of interest (POI), in terms of total number of shops and restaurants within a 500-meter radius from each metro station, indicated by the depth of colour in Figure 2. This variable is regarded as a proxy for the attractiveness of each station for after-work activities in this study. It can be observed that the spatial distribution of POIs is concentrated towards the central part of the city, and it is also interesting to notice that in distant areas from the city centre, that distribution is concentrated in one or two stations, which can be interpreted as being city sub-centres.

3.2 Smart card data

One of the ways in which the government promoted PT in Shanghai was to introduce the automated fare collection system that automates the ticketing system for the entire PT network, including metro, bus, taxi, ferry and P+R. Travelers are allowed to pay these services by using a smart card not only for its convenience but also to get a discount.

The SCD provided by the SODA contest contains the records of all transactions by all smart cards in April, 2015. In Shanghai, metro is the only PT system where card holders should both check in and check out. On the other hand, travellers are required to scan their cards only when boarding a bus or alighting a taxi, not to mention that the location information is missing on these modes. Therefore, we focused on the metro network for further analysis and modelling.

In addition, we carefully dealt with those trips including transfers. In Shanghai, only a few metro stations require travellers to check out and then check in again to switch to another line. Such cases should not be seen as two separate trips. To distinguish them, we used a threshold of 30 minutes between check-out and check-in at those stations. The selection of this threshold is based on the policy by which after 30 minutes without checking in again, the system will regard the next check-in as the start of a new trip. We assume that travellers are aware of this fact, and if they stay at those stations for more than 30 minutes, they must have performed an activity whose utility can compensate for the loss.

4. Results of the case study

4.1 Detecting metro commuters

After applying the method for detecting the commuters, there were about 0.8 million metro commuters filtered from the data. This number can be compared with the average daily number of unique card IDs scanned for metro trips, which was about 2 million. We did not include those commuters who had detected PT access and/or egress modes such as bus trips connecting with metro trips for commuting. Figure 3 shows the spatial distributions of home stations and workplace stations of all the detected metro commuters. By comparing the spatial distributions of home stations, workplace stations and POIs (shown in Figure 2 and Figure 3), we found that the spatial distribution of home stations was completely different from the ones of workplace stations and POIs, and the latter two were somehow similar to each other.

[2] http://dianping.com/ (retrieved date: November 21st, 2016)

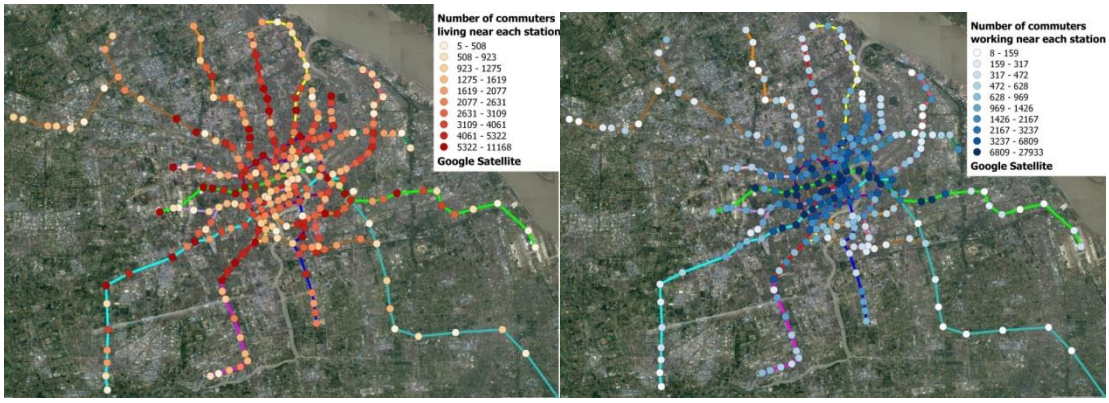


Figure 3 Number of metro commuters living near each station (left) and number of metro commuters working near each station (right)

4.2 Extracting daily metro trip chains

In our study, we focused on the metro commuters and extracted their daily metro trip chains which only consisted of metro trips. The ten most common types of the daily metro trip chains are plotted in Figure 4. Among the metro commuters on an average weekday, about 64.7% performed the home-work-home chain, which was the most common type of trip chains, and at least 13.5% performed the trip chains involving secondary activities. This shows that neglecting this kind of travel patterns may cause the distortion of travel demand prediction.

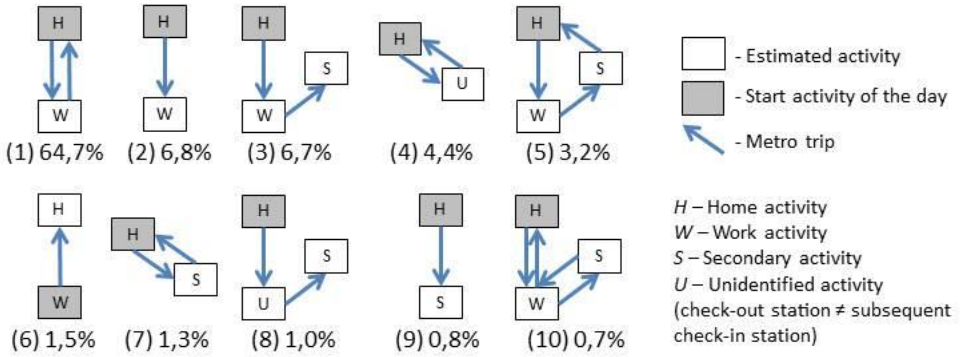


Figure 4 The top 10 most common types of daily metro trip chains

For chain type (3), (5) and (7) to (10), which involve secondary activities, the mean arrival-departure hours of the day at the secondary activity are 18-NA (no departure), 18-20, 10-17, 18-NA (no departure), 10-NA (no departure) and 13-14 respectively. It seems that type (10) is more likely to indicate a person who has a lunch break from work, and type (7) and (9) correspond more to business trips. The other three types are more related to the travel patterns of an individual performing an after-work activity.

4.3 Model estimation

We focused on the after-work activities which were performed after 16:00 in chain type (5). Considering the computational limits, we randomly selected 3,000 commuters who experienced the prescriptive choice situations in the month. To explain the revealed station choice behaviour, we used the model structure proposed in Section 2.3. The variable

specifications in the utility function formulated as Equation (3) should be updated in the context of the case study. The attractiveness of a station for after-work activities was defined as the number of POIs around the station. The features of travel impedance included metro network distance and number of metro transfers. As the characteristics of an after-work activity, activity duration was assumed to be the time gap between the arrival time and the departure time at the station for an after-work activity, and activity start time was quantified by the time gap between 16:00 and the arrival time at the station for the after-work activity.

We have calculated the spatial distribution of home and workplace stations of all the metro commuters (See Figure 3). Based on that, we can calculate for each station the ratio of the number of commuters living there over the number of commuters working there, and this ratio is designated as the residents-to-jobs ratio (RJ ratio). For each commuter, we further calculated the RJ ratios for the home station and the workplace station respectively. It can be observed in Figure 3 that if the RJ ratio of one's home station is higher, then this person is more likely to live in a mainly residential area, located in the peripheral area of Shanghai; Otherwise, this person is more likely to live in a mainly commercial area, located in the central area of Shanghai. The same applies to interpreting the RJ ratio of one's workplace station. These two variables, along with the commuting network distance and the number of transfers along the commuting trip, can serve as proxies for some personal distinction among the travellers. For each choice occasion, we computed the choice set defined by the constraints specified in Section 2.3. Figure 5 shows the distribution of the size of the choice set, in terms of the number of reachable stations divided by the total number of stations in the network excluding the home and workplace. In about 78% of the choice occasions, there is a set of stations that a traveller cannot choose due to the constraints.

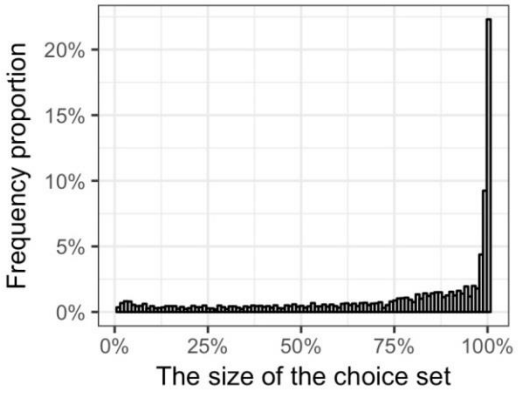


Figure 5 The histogram of the size of the choice set

The estimation results are compared under different model specifications. First, we tested how the different ways of defining travel impedance (See Table 1) would influence model fit. Second, we tested how the introduction of the last choice feedback variable would lead to different model estimates.

The estimation results of the models using different travel impedances without considering last choice feedback are presented in Table 4, Table 5 and

Table 6, where only the statistically significant estimates are retained (p-value < 0.05). Biogeme (Bierlaire, 2003) is the software package we used for model estimation in this study.

Table 4 The estimation results of the discrete choice model using home-based travel impedance without considering last choice feedback

Variable	Parameter	Robust t-test value
Number of POIs	**4.28e-04**	**10.00**
Number of POIs × activity duration	3.98e-05	6.06
Number of POIs × commuting network distance	3.32e-05	2.43
Number of POIs × RJ ratio of home station	1.53e-05	2.51
Number of POIs × RJ ratio of workplace station	-1.10e-05	-2.27
Network distance from home	**-0.255**	**-12.6**
Network distance from home × activity duration	0.0140	4.86
Network distance from home × activity start time	-0.00987	-3.19
Network distance from home × commuting network distance	0.0592	9.14
Network distance from home × commuting number of transfers	-0.0223	-4.25
Number of transfers from home	**-0.848**	**-4.62**
Number of transfers from home × activity duration	0.184	6.45
Number of transfers from home × commuting network distance	-0.293	-4.89
Number of transfers from home × commuting number of transfers	0.544	9.63
Number of transfers from home × RJ ratio of home station	0.0616	2.38
Number of transfers from home × RJ ratio of workplace station	0.0702	3.29
Number of observations: 5107; Initial log likelihood: -26589.161; Final log likelihood: -21128.360; Adjusted rho-square: 0.205; Run time: 1’ 58’’		

Table 5 The estimation results of the discrete choice model using detour travel impedance without considering last choice feedback

Variable	Parameter	Robust t-test value
Number of POIs	**4.44e-04**	**9.31**
Number of POIs × activity duration	6.80e-05	9.13
Number of POIs × RJ ratio of home station	1.59e-05	2.34
Number of POIs × RJ ratio of workplace station	3.03e-05	5.58
Detour network distance	**-0.0579**	**-3.60**
Detour network distance × activity duration	0.00915	3.26
Detour network distance × commuting network distance	-0.0174	-3.43
Detour network distance × commuting number of transfers	-0.0134	-2.17
Detour number of transfers	**-0.918**	**-7.58**
Detour number of transfers × activity duration	0.180	8.98
Detour number of transfers × commuting network distance	-0.0795	-2.08
Detour number of transfers × RJ ratio of workplace station	0.0840	5.41
Number of observations: 5107; Initial log likelihood: -26589.161; Final log likelihood: -20530.100; Adjusted rho-square: 0.227; Run time: 1’46’’		

Table 6 The estimation results of the discrete choice model using detour travel impedance and proximity to home vs. workplace without considering last choice feedback

Variable	Param.	Robust t-test value
Number of POIs	**4.21e-04**	**8.73**
Number of POIs × activity duration	5.50e-05	7.30
Number of POIs × RJ ratio of home station	2.00e-05	2.94
Number of POIs × RJ ratio of workplace station	2.26e-05	3.98
Detour network distance	**-0.0613**	**-3.84**
Detour network distance × activity duration	0.00688	2.47
Detour network distance × commuting network distance	-0.0152	-3.01
Detour network distance × commuting number of transfers	-0.0121	-2.00
Detour network distance × RJ ratio of home station	0.00589	2.24
Detour number of transfers	**-0.931**	**-7.59**
Detour number of transfers × activity duration	0.171	8.47
Detour number of transfers × RJ ratio of workplace station	0.0802	5.10
Home vs. workplace proximity in terms of network distance (Network distance from home – network distance from workplace)	**-0.0676**	**-3.78**
Home vs. workplace proximity (network distance) × activity duration	0.0131	5.32
Home vs. workplace proximity (network distance) × activity start time	-0.00717	-3.00
Home vs. workplace proximity (network distance) × commuting network distance	0.0185	3.22
Home vs. workplace proximity in terms of number of transfers (Number of transfers from home – number of transfers from workplace)	**0.414**	**2.74**
Home vs. workplace proximity (number of transfers) × activity start time	-0.0863	-3.12
Home vs. workplace proximity (number of transfers) × commuting network distance	-0.113	-2.29
Number of observations: 5107; Initial log likelihood: -26589.161; Final log likelihood: -20404.619; Adjusted rho-square: 0.231; Run time: 3’ 10’’		

First, the effects of spatial autocorrelation are found to be statistically insignificant in all cases as the estimated values of the dissimilarity parameter ρ are not significantly different from 1. Thus, the spatially correlated model structure actually collapses to the multinomial logit one, of which we present the results. Second, we see that the metro commuters significantly prefer to visit the stations where there are more POIs for performing after-work activities, which is not a surprise. Third, the model using both detour impedance and home vs. workplace proximity fits the data slightly better than the model using only detour impedance, and both of them outperform the one using home-based impedance. This result substantiates the research conclusion drawn by Arentze and Timmermans (2007) regarding the benefit of modelling detour travel impedance, and apart from that, it further shows that commuters do give different weights to travel impedance to access an after-work activity coming from home or from the workplace. It turns out that they generally prefer the stations which are closer from the workplace in terms of number of transfers but closer from home in terms of network distance, ceteris paribus. Fourth, the attributes related to activities are observed to have a considerable impact on station choices for after-work

activities. The results significantly show that people give a higher weight to the number of POIs and care less about all kinds of travel impedances if the activity duration is longer. In addition, an activity of longer duration is preferred to take place near the workplace station than near the home station in terms of network distance. The activity start time is an especially effective variable interacting with the home vs. workplace proximity. It can be observed that for a later activity, people's preference for reducing travel impedance from home weighs more than reducing the one from workplace. Fifth, results seem to support the use of proxy variables to translate differences between travellers. Given that an individual has longer commuting distance, this person seems to be more reluctant to detour farther for after-work activities. A commuter whose home station has higher RJ ratio is more willing to visit a station with a greater number of POIs for after-work activities.

We also estimated the model using detour travel impedance and home vs. workplace proximity after considering last choice feedback. The first choice of each traveller was not modelled since it was assumed to be exogenously given. The estimation results are shown in Table 7.

Table 7 The estimation results of the discrete choice model using detour travel impedance and home vs. workplace proximity considering last choice feedback

Variable	Param.	Robust t-test value
Number of POIs	**3.88e-04**	**2.84**
Number of POIs × activity duration	7.15e-05	3.28
Number of POIs × commuting network distance	-1.06e-04	-2.81
Detour network distance	**-0.0734**	**-2.84**
Detour network distance × activity duration	0.0107	2.61
Detour network distance × commuting network distance	-0.0210	-2.94
Detour number of transfers	**-0.851**	**-3.63**
Detour number of transfers × activity duration	0.141	4.26
Detour number of transfers × RJ ratio of workplace station	0.0872	3.30
Home vs. workplace proximity in terms of network distance	**-0.144**	**-6.14**
Home vs. workplace proximity (network distance) × activity duration	0.0108	3.22
Home vs. workplace proximity (network distance) × commuting network distance	0.0439	6.24
Home vs. workplace proximity in terms of number of transfers	**0.689**	**2.24**
Home vs. workplace proximity (number of transfers) × commuting network distance	-0.244	-2.69
Last choice feedback	**3.99**	**60.39**
Number of observations: 2127; Initial log likelihood: -11448.617; Final log likelihood: -6378.684; Adjusted rho-square: 0.440; Run time: 2' 38''		

Again the effect of spatial correlation is not statistically significant in this model. It can be observed that travellers frequently chose the same station for after-work activities, leading to the overwhelmingly significant estimate of the preference for the last choice feedback variable which leads to a better model fit. Such a good fit does not necessarily lead to a good demand prediction in future scenarios, because the model relies heavily on the assumption that the previous choice is exogenously given. However, this model can still help us figure out whether we misestimate any parameters due to neglecting habitual effect. After

introducing the variable of last choice feedback, results indicate that travellers actually do not give as much weight to the number of POIs as was estimated previously. To make a choice among those stations which have not been visited previously, people seem to care less about detour number of transfers but care more about detour network distance, and they are more likely to choose a station even closer to home in terms of network distance. The effect of activity start time is no longer significant on the preference for home vs. workplace impedance, indicating that this effect estimated in the previous models might have been related with habitual behaviour. However, the effects of activity duration and commuting network distance on the preferences still exist.

5. Conclusions and recommendations

In this paper, after detecting metro commuters and extracting their trip chains from the SCD, we focused on modelling their station choices for after-work activities. The method was applied to the case study of metro travellers in Shanghai. The advantages of using SCD over travel survey data for this purpose include the cost efficiency of data collection, the full population of travellers, and the revealed panel effect. In addition, to overcome the drawback of such anonymous data, we proposed to use proxy variables to distinguish the travellers, which can help better explain the heterogeneity of location choice behaviour among the population. Moreover, different ways of modelling travel impedance were compared, and we found that the model using detour impedance and home vs. workplace proximity, which we created in this study to model the travel impedance to conduct after-work activities, outperformed the others and improved the interpretation of behaviour.

This work can still be improved in a few ways. First, a travel survey dataset is recommended to be complementarily used for validation and reference. It can help improve the accuracy of commuter detection and identify more specific activity purposes among after-work activities. Also, stated-preference data from travel survey can potentially help enhance the understanding of how travellers perceive travel impedance for after-work activities, further improving our proposed travel impedance metrics. For example, the preference for reducing travel impedance may be related to factors such as familiarity with a particular area, which is difficult to obtain using smart card data. Next, the discrete choice model can be further elaborated to take more factors into consideration. Finally, we only focused on the station choices for after-work activities conducted in a certain type of daily trip chain in this study; however, a more general framework can be built to model station choices for all secondary activities using SCD in future research.

Acknowledgement

We would like to express our gratitude to the Shanghai Open Data Apps (SODA) contest for making the data available for this research. Thanks go also to the TRAIL research school and the Dutch Organization for Scientific Research (NWO) for sponsoring the first author for his PhD study.

References

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin – destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C 58, 240–250. doi:10.1016/j.trc.2015.02.018

Arentze, T.A., Timmermans, H.J.., 2004. A learning-based transportation oriented simulation system. Transp. Res. Part B Methodol. 38, 613–633. doi:10.1016/j.trb.2002.10.001

Arentze, T., Timmermans, H., 2007. Robust Approach to Modeling Choice of Locations in Daily Activity Sequences. Transp. Res. Rec. J. Transp. Res. Board 2003, 59–63. doi:10.3141/2003-08

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transp. Policy 12, 464–474. doi:10.1016/j.tranpol.2005.06.008

Balmer, M., Meister, K., Nagel, K., 2008. Agent-based simulation of travel demand: Structure and computational performance of MATSim-T.

Bhat, C.R., Guo, J., 2004. A mixed spatially correlated logit model: formulation and application to residential choice modeling. Transp. Res. Part B Methodol. 38, 147–168. doi:10.1016/S0191-2615(03)00005-5

Bierlaire, M., 2003. BIOGEME: a free package for the estimation of discrete choice models. Proc. 3rd Swiss Transp. Res. Conf.

Castillo-Manzano, J.I., López-Valpuesta, L., 2009. Urban retail fabric and the metro: A complex relationship. Lessons from middle-sized Spanish cities. Cities 26, 141–147. doi:10.1016/j.cities.2009.02.007

Chakirov, A., Erath, A., 2012. Activity Identification and Primary Location Modelling based on Smart Card Payment Data for Public Transport Smart Card Payment Data for Public Transport. Int. Conf. Travel Behav. Res.

Danalet, A., Tinguely, L., Lapparent, M. de, Bierlaire, M., 2016. Location choice with longitudinal WiFi data. J. Choice Model. 18, 1–17. doi:10.1016/j.jocm.2016.04.003

De Vos, J., Schwanen, T., Van Acker, V., Witlox, F., 2013. Travel and Subjective Well-Being: A Focus on Findings, Methods and Future Research Needs. Transp. Rev. 33, 421–442. doi:10.1080/01441647.2013.815665

Demerouti, E., Bakker, A.B., Geurts, S.A.E., Taris, T.W., 2009. Daily recovery from work-related effort during non-work time. Curr. Perspect. Job-Stress Recover. Res. Occup. Stress Well Being 7, 85–123. doi:10.1108/S1479-3555(2009)0000007006

Devillaine, F., Munizaga, M., Trépanier, M., 2013. Detection of Activities of Public Transport Users by Analyzing Smart Card Data. Transp. Res. Rec. J. Transp. Res. Board 48–55. doi:10.3141/2276-06

Dong, X., Ben-akiva, M.E., Bowman, J.L., Walker, J.L., 2006. Moving from trip-based to activity-based measures of accessibility 40, 163–180. doi:10.1016/j.tra.2005.05.002

Goh, S., Lee, K., Park, J.S., Choi, M.Y., 2012. Modification of the gravity model and application to the metropolitan Seoul subway system. Phys. Rev. E 86, 26102. doi:10.1103/PhysRevE.86.026102

Goulet Langlois, G., Koutsopoulos, H.N., Zhao, J., 2016. Inferring patterns in the multi-week activity sequences of public transport users. Transp. Res. Part C Emerg. Technol. 64, 1–16. doi:10.1016/j.trc.2015.12.012

Horni, A., 2013. Destination choice modeling of discretionary activities in transport microsimulations. doi:10.3929/ETHZ-A-010006641

Ibrahim, M.F., McGoldrick, P.J., 2003. Shopping choices with public transport options : an agenda for the 21st century. Ashgate.

Koppelman, F.S., Wen, C.-H., 2000. The paired combinatorial logit model: properties, estimation and application. Transp. Res. Part B Methodol. 34, 75–89. doi:10.1016/S0191-2615(99)00012-0

Kuhlman, W., 2015. The construction of purpose-specific OD matrices using public transport smart card data. Delft University of Technology.

Liu, L., Hou, A., Biderman, A., Ratti, C., Chen, J., 2009. Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. Proc. 2009 12th Int. IEEE Conf. Intell. Transp. Syst. 1–6. doi:10.1109/ITSC.2009.5309662

Long, Y., Thill, J.-C., 2015. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. Comput. Environ. Urban Syst. 53, 19–35. doi:10.1016/j.compenvurbsys.2015.02.005

Long, Y., Zhang, Y., Cui, C., 2012. Analysing jobs-housing relationship and commuting patterns of Beijing using bus smart card data (in Chinese). Acta Geogr. 67, 1339–1352. doi:10.11821/XB201210005

Lu, X., Gu, X., 2011. The Fifth Travel Survey of Residents in Shanghai and Characteristics Analysis. Urban Transp. China 9, 1–7.

Ma, X., Liu, C., Wen, H., Wang, Y., Wu, Y.-J., 2017. Understanding commuting patterns using transit smart card data. J. Transp. Geogr. 58, 135–145. doi:10.1016/j.jtrangeo.2016.12.001

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transp. Res. Part C Emerg. Technol. 36, 1–12. doi:10.1016/j.trc.2013.07.010

Morency, C., Trépanier, M., Agard, B., 2007. Measuring transit use variability with smart-card data. Transp. Policy 14, 193–203. doi:10.1016/j.tranpol.2007.01.001

Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. Transp. Res. Part C Emerg. Technol. 24, 9–18. doi:10.1016/j.trc.2012.01.007

Nakamura, K., Gu, F., Wasumtarasook, V., Vichiensan, V., Hayashi, Y., 2016. Failure of Transit-Oriented Development in Bangkok from a Quality of Life Perspective. Asian Transp. Stud. 4, 194–209. doi:10.11175/EASTSATS.4.194

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transp. Res. Part C Emerg. Technol. 19, 557–568. doi:10.1016/j.trc.2010.12.003

Rasouli, S., Timmermans, H., 2014. Activity-based models of travel demand: promises, progress and prospects. Int. J. Urban Sci. 18, 31–60. doi:http://dx.doi.org/10.1080/12265934.2013.835118

Schönfelder, S., Axhausen, K.W., 2003. Activity spaces: measures of social exclusion? Transp. Policy 10, 273–286. doi:10.1016/j.tranpol.2003.07.002

Sivakumar, A., Bhat, C., 2007. Comprehensive, unified framework for analyzing spatial location choice. Transp. Res. Rec. J. Transp. Res. Board 103–111. doi:10.3141/2003-13

Sun, L., Lu, Y., Jin, J.G., Lee, D.-H., Axhausen, K.W., 2015. An integrated Bayesian approach for passenger flow assignment in metro networks. Transp. Res. Part C Emerg. Technol. 52, 116–131. doi:10.1016/j.trc.2015.01.001

Suriñach, J., Romaní, J., Royuela, V., Reyes, M., 2000. Urban systems in the Barcelona province: A first step for estimating local economic activity, in: Paper for the 40th European Congress of the European Regional Science Association, Barcelona.

Train, K., 2009. Discrete choice methods with simulation. Cambridge University Press.

Trépanier, M., Tranchant, N., Chapleau, R., 2007. Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. J. Intell. Transp. Syst.

Zhou, J., Murphy, E., Long, Y., 2014. Commuting efficiency in the Beijing metropolitan area: an exploration combining smartcard and travel survey data. J. Transp. Geogr. 41, 175–183. doi:10.1016/j.jtrangeo.2014.09.006