Journal of Transport Geography 63 (2017) 40–47

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

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, The Hague, 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

ARTICLE INFO

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

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.

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 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

* Corresponding author.
E-mail address: y.wang-14@tudelft.nl (Y. Wang).

http://dx.doi.org/10.1016/j.jtrangeo.2017.06.010
Received 2 January 2017; Received in revised form 19 May 2017; Accepted 30 June 2017
0966-6923/

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.

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

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-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 h 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 and secondary activity dependent on whether the station is the home or workplace station 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