Transportation Research Part C 92 (2018) 76–89

Contents lists available at ScienceDirect

Transportation Research Part C

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

Estimating metro passengers' path choices by combining self-reported revealed preference and smart card data

Yongsheng Zhang[a], Enjian Yao[a,*], Junyi Zhang[b], Kangning Zheng[a]

[a] School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China
[b] Graduate School for International Development and Cooperation, Hiroshima University, 1-5-1 Kagamiyama, Higashi Hiroshima 739-8529, Japan

ARTICLE INFO

Keywords:
Data fusion
Expectation-Maximization algorithm
Metro network
Stochastic travel time budget
Risk-averse attitude

ABSTRACT

With the help of automated fare collection systems in the metro network, more and more smart card (SC) data has been widely accumulated, which includes abundant information (i.e., Big Data). However, its inability to record passengers' transfer information and factors affecting passengers' travel behaviors (e.g., socio-demographics) limits further potential applications. In contrast, self-reported Revealed Preference (RP) data can be collected via questionnaire surveys to include those factors; however, its sample size is usually very small in comparison to SC data. The purpose of this study is to propose a new set of approaches of estimating metro passengers' path choices by combining self-reported RP and SC data. These approaches have the following attractive features. The most important feature is to jointly estimate these two data sets based on a nested model structure with a balance parameter by accommodating different scales of the two data sets. The second feature is that a path choice model is built to incorporate stochastic travel time budget and latent individual risk-averse attitude toward travel time variations, where the former is derived from the latter and the latter is further represented based on a latent variable model with observed individual socio-demographics. The third feature is that an algorithm of combining the two types of data is developed by integrating an Expectation-Maximization algorithm and a nested logit model estimation method. The above-proposed approaches are examined based on data from Guangzhou Metro, China. The results show the superiority of combined data over single data source in terms of both estimation and forecasting performance.

1. Introduction

With the development of metro networks, for example in Beijing, Shanghai, Guangzhou, Tokyo, London, New York City and Singapore, more and more metro passengers can choose their most preferred traveling path from multiple transit paths. This has enhanced the convenience and attractiveness of metro systems on one hand, while it has become a challenge how to capture such multiple path choices for travel demand prediction and management, as well as ticket fare allocation among different transit operators. Even though the above multiple path choices have been observed in many years ago, re-visiting this issue has its contemporary implications, especially considering the progress of information and communication technologies, which have brought various opportunities and challenges to transit network planning and management.

Usually, the above path choice problem is solved by using path choice models. Given a path choice model (e.g., a multinomial logit (MNL) model derived from the random utility maximization (RUM) principle), the most important thing is to collect enough choice data for estimating it. Traditionally, Revealed Preference (RP) and Stated Preference (SP) data are the two major sources

* Corresponding author.
E-mail addresses: 12114241@bjtu.edu.cn (Y. Zhang), enjyao@bjtu.edu.cn (E. Yao), zjy@hiroshima-u.ac.jp (J. Zhang), 15120932@bjtu.edu.cn (K. Zheng).

https://doi.org/10.1016/j.trc.2018.04.019
Received 28 October 2017; Received in revised form 21 April 2018; Accepted 21 April 2018
Available online 03 May 2018
0968-090X/

Table 1
The major differences between self-reported RP and SC data.

	Self-reported RP data	SC data
Travel path	Yes	No
Individual socio-demographics	Yes	No
Exact travel time	No	Yes
Sample size	Limited via sampling methods	Close to the whole population
Collection manner	Questionnaire surveys	AFC systems

widely used in the transportation literature and practices (e.g., Morikawa, 1989). RP data reveals people's actual travel behaviors, while SP data collects people's travel behaviors under hypothetical choice scenarios. To mitigate the influences of various biases in SP data, data fusion techniques have been proposed to combine RP data with SP data for capturing people's true preferences in a more accurate way (e.g., Ben-Akiva and Morikawa, 1990). More choice data can be collected via SP surveys in a relatively efficient way than RP surveys, because SP surveys can collect two or more choices from a single respondent. Nevertheless, it is costly and time-consuming for both types of surveys to collect large-scale data.

In addition to the above two types of questionnaire surveys, RP data can be obtained via various technologies, such as GPS (e.g., Zimmermann et al., 2017), and mobile phones (e.g., Wang et al., 2018), in a passive way (also called passive data). Hereafter, RP data collected from questionnaire surveys is named as 'self-reported RP data' in this study, because questionnaire surveys usually allow respondents to report the travel behavior related information by themselves. However, the above RP data collected with the assistance of technologies suffers from privacy issues and high equipment costs in practice. In the context of metro networks, Automated Fare Collection (AFC) systems, which have been developed for fare charges, have collected and stored huge amount of the so-called 'smart card (SC) data'. Such SC data records each traveler's actual travel information, including entry station, entry time, exit station, exit time, and ticket fare, but it does not include travelers' socio-demographics and the detailed information of travel paths. It should be noted that for some metro systems, such as New York City Subway, only the tap-in information has been collected.

This study only focuses on the SC data containing both entry and exit information, which can be collected in all Chinese cities with metro systems, where a Chinese city is targeted. As for metro operation, this type of data has become an attractive data source for analyzing and predicting passengers' travel characteristics (e.g., actual travel times, departure times, travel choice preferences) and flow distributions within the metro networks.

Table 1 summarizes the major differences of self-reported RP and SC data. Self-reported RP data records passenger's actual travel paths on some specific days and some individual-level socio-demographics (e.g., age, gender, income, and purpose). It can collect travel time, which is however not accurate, and furthermore, the sample size is small. In contrast, SC data covers all passengers and records each passenger's OD pairs and travel time over the whole operation period of every day, which can be used to capture travel time uncertainty caused by recurrent congestion (e.g., passengers crowding in the platforms and walking corridors) and non-recurrent congestion (e.g., incidents during transit operation). However, actual travel paths and individual-level socio-demographics are unknown in SC data. Obviously, the two types of data have complementary advantages. Therefore, it is straightforward and meaningful to combine the two types of data for better serving transportation planning and management.

In short, this study will propose a set of approaches to estimate metro passengers' path choices by combining self-reported RP and SC data, where the effects of travel time uncertainty will be reflected.

The remaining part of this study first reviews existing studies and describes the path choice model with combined self-reported RP and SC data. After that, it explains a new algorithm developed for estimating the model. Furthermore, it presents numerical analysis results. Finally, it concludes this study, together with a discussion about future research issues.

2. Literature review

RUM-based models (e.g., McFadden, 1968; Daganzo and Sheffi, 1977) have been widely used in travel behavior analysis, including path choice modeling. Compared with RUM-based probit models (Daganzo and Sheffi, 1977), RUM-based logit models have been applied more widely because of their closed forms of probability equations and ease of estimation and application. Among various logit-type models, MNL model (Dial, 1971) is the most operational choice model. However, MNL model suffers from the independence of irrelevant alternatives (IIA) property, because error terms are assumed to follow an identical and independent distribution (IID). To overcome this shortcoming, many alternative logit-type models have been developed, such as C-logit model (Cascetta et al., 1996), Cross-nested Logit (CNL) model (Vovsha, 1997), Implicit Availability/Perception (IAP) model (Cascetta et al., 1999), Path-sized Logit (PSL) model (Ben-Akiva and Bierlaire, 1999), Paired Combinatorial Logit (PCL) model (Koppelman and Wen, 2000), Generalized Nested Logit (GNL) model (Wen and Koppelman, 2001), Mixed Logit model (McFadden and Train, 2000), and Recursive Logit (RL) model (Fosgerau et al., 2013).

Travel behavior analysis has applied RP and SP data, mainly since 1980s (e.g. Morikawa, 1989). Even in recent years, one can still find many studies applying them to examine, for example, travel time reliability (e.g., Swierstra et al., 2017), route choice (e.g., Yang et al., 2016), destination choice (e.g., Clifton et al., 2016), and mode choice (e.g., R. Zhang et al., 2017). SP data usually collects two or more choice samples from a respondent based on hypothetical choice scenarios, while RP data can collect only one choice sample per respondent. Thus, SP data is more efficient in terms of data collection than RP data. However, SP involves more biases than RP data. For mitigating the influence of biases in SP data on model estimation, many studies have been done to combine these two types