Transportation Research Part C 44 (2014) 70–79

Contents lists available at ScienceDirect

Transportation Research Part C

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

# Validating travel behavior estimated from smartcard data

Marcela Munizaga *, Flavio Devillaine, Claudio Navarrete, Diego Silva

*Departamento de Ingeniería Civil, Universidad de Chile, Blanco Encalada 2002, Santiago, Chile*

ARTICLE INFO

*Article history:*
Received 11 December 2013
Received in revised form 11 March 2014
Accepted 12 March 2014

*Keywords:*
Validation
Smartcard data
OD matrices
Public transport

ABSTRACT

In this paper, we present a validation of public transport origin–destination (OD) matrices obtained from smartcard and GPS data. These matrices are very valuable for management and planning but have not been validated until now. In this work, we verify the assumptions and results of the method using three sources of information: the same database used to make the estimations, a Metro OD survey in which the card numbers are registered for a group of users, and a sample of volunteers. The results are very positive, as the percentages of correct estimation are approximately 90% in all cases.



## 1. Introduction

The acquisition of travel information from smartcard data is a growing trend. Many researchers have foreseen the opportunity to obtain high-quality information at a very low cost and have therefore developed tools to obtain valuable information generated as a side-product from original data on the operation of transit systems. Bagchi and White (2005) and Utsunomiya et al. (2006) presented visionary discussions of the potential of smartcard data. Pelletier et al. (2011) performed a complete review of the different applications explored to date. Among the different types of applications reported and envisioned, the use of rules-based processing to obtain OD matrices is a recurrent practice. Some examples of this application include those described by Barry et al. (2002), who used MetroCard data from New York to estimate station-to-station OD flows; Lianfu et al. (2007), who proposed a numerical method to build an OD matrix at the bus-stop level, using data from Changchun, China; Zhao et al. (2007), who inferred rail passenger trip OD matrices from an origin-only automatic fare collection system, in which the position of buses is known due to an automatic vehicle location system; Trépanier et al. (2007), who focused on the destination estimation method and applied it to data from Gatinueau, Canada; Wang et al. (2011), who developed an application for Oyster card data from London; and Munizaga and Palma (2012), who proposed modifications to previous methods to adapt them to the large and complex network of Transantiago (Chile).

In the majority of these cases, the estimation of alighting stops is crucial to the reliability of the results obtained, but many other assumptions have to be made to obtain valuable results. However, most authors recognize that the assumptions made in the process of estimating destination, route and activities from raw smartcard data need to be validated. A few attempts at validation have been made. For example, Barry et al. (2009) proposed validating the data obtained from smartcards by conducting a comparison with other sources of information, such as exit and entrance counts at subway stations and bus ride check data (counts of boarding and alighting flows as well as overall loads). However, additional information is not always available with the coverage and quality required to conduct a validation process. Farzin (2008) compared the aggregate results of an origin–destination (OD) matrix obtained from approximately 5% of the total trips (due to a lack of GPS

* Corresponding author. Tel.: +56 229784649
*E-mail address:* mamuniza@ing.uchile.cl (M. Munizaga).

http://dx.doi.org/10.1016/j.trc.2014.03.008
0968-090X/

equipment on many buses) with the results from an OD survey conducted almost 10 years before the smartcard data were obtained. The results were not conclusive due to the lack of representativeness for both matrices. Wang et al. (2011) also compared data with boarding/alighting counts and obtained positive results at an aggregate level. Devillaine et al. (2013) analyzed the potential bias of an OD matrix obtained from smartcard data and, with that focus, explored the possibility of endogenous validation, analyzing the available data and the results of the methods applied in processing the data. The authors also attempted to validate their results with exogenous data, but the survey that they utilized corresponded to a different week than the available smartcard data.

The objective of this paper is to propose and apply a series of methods to validate the assumptions made in the set of methodologies used to input boarding positions, alighting stops, routes chosen, activities at the destination and purpose assignment for the case of Transantiago (a public transport system in Santiago, Chile). An OD matrix was constructed using approximately 80% of the boarding transactions, including over 20 million observed trips in a week (Munizaga and Palma, 2012; Devillaine et al., 2012). Three sources of information are available for validation:

- information from the same database used to make the estimations (endogenous validation);
- information from a detailed origin–destination survey applied to a sample of 300,000 Metro users, of which a small percentage provided their smartcard ID for validation purposes; and
- personal interviews from a small sample of volunteers who provided travel and personal information.

Analyses of the information used to make the estimations and of the estimations themselves (endogenous validation) can be useful to detect errors and thus improve the methodology. The large sample of OD surveys contains information that can be used to compare the route chosen by the users within the Metro network with the route assigned by the model (minimum cost). The small sample of users who provided their card ID number can be used to explore the stronger and most crucial assumption of the model: the estimation of alighting stop. The sample of volunteers also allows for the exploration of more ambitious estimation such as the identification of trip destinations/transfers and trip purpose. In the remainder of this section, we synthesize the most relevant aspects of the methodology to be validated (details can be found in the cited references). Section 2 describes the endogenous validation, Section 3 describes the exogenous validation with the Metro OD survey data, Section 4 describes the validation based on volunteers and Section 5 concludes.

### *1.1. Description of the methodology to be validated*

Munizaga and Palma (2012) proposed a method for observing card transactions in public transport systems and for estimating travel sequences using information from transaction sequences. Only boarding transactions were observed because the payment system does not require validation when alighting. The assumptions that Munizaga and Palma (2012) made to estimate a trip matrix were as follows:

- Trip stages begin at the time/location when/where the validations occur.
- The end of a trip stage can be found at the stop or station most convenient to reach the next boarding location. This station would be the nearest in the case of the Metro and the stop that minimizes the generalized time (weighted function of vehicle travel time and estimated walking time) for bus and bus station transactions. In the last case, common lines are considered to identify possible routes. In all cases, only stops within walking distance (1 km) are considered. Within the Metro network, a deterministic route choice (minimum travel time) is assumed.
- Trips are defined as sequences of trip stages with less than 30 min between the end of one stage and the beginning of the next, without consecutive validations in the Metro or on the same bus route.

Using these assumptions, the alighting stop was estimated for over 80% of the boarding transactions, and OD matrices were built. The resulting OD matrices appear reasonable, being much denser than the OD matrices obtained from surveys. However, Munizaga and Palma (2012) recognized that more sophisticated methods for identifying trips and trip stages were required. The correct identification of trips and trip stages is crucial to obtain reliable OD matrices. Origins and destinations are locations at which the needs of the users are to be satisfied through engagement in activities, while transfers are simply a consequence of the interaction between the transit network and those needs. Using the results of Munizaga and Palma (2012), Devillaine et al. (2012) proposed a method for estimating the location, duration and purpose of activities. Given a daily sequence of transactions, the time intervals between estimated trips are regarded as work if the time elapsed between the estimated alighting from one trip and the time of boarding for the next is more than 5 h in the case of regular card users. For student cards, if the time elapsed between the estimated alighting and the next boarding is more than 2 h, study activity is assumed for that lapse. If the time elapsed between the estimated alighting and the next boarding was less than 2 h, we assumed an activity categorized as "Other" both for regular users and students. Home was assumed to be the destination after the estimated alighting for the last transaction of the day and the first boarding transaction of the next day. Following the analysis of Devillaine et al. (2013) regarding the validity of these assumptions, we explore the use of endogenous validation and propose new rules in Section 2. These new rules are tested using the exogenous validation sample in Section 4.