Title:

Comparing origin-destination matrices derived from smart card data with a large scale OD survey. Results from a case study in Lyon.

Author(s):

Oscar EGU,
Keolis Lyon, 19 Boulevard Marius Vivier Merle, 69003 Lyon
oegu@keolis-lyon.fr

Patrick Bonnel,
LAET - ENTPE, Rue Maurice Audin, 69518 Vaulx-en-Velin Cedex,
Patrick.Bonnel@entpe.fr

Keywords :

Public transportation, big data, smart card data, survey, OD matrices

Short abstract:

Origin-destination matrices are essential to understand movement within a city and public transit demand. Those matrices are used by public transport operators to estimate load, to evaluate the impact of networks modifications but also to understand passenger’s behaviors. Traditionally the data needed to build such matrices were collected using survey. However, they remain costly and are limited in terms of spatiotemporal coverage. At the same time, the recent advance in data collection and the large spread of automatic fare collection has made it possible to build origin-destination matrices using only passive data stream. Those new data sources are quite appealing and could replace the traditional method on the condition that they provide results of similar quality. The main objective of this research is, therefore, to compare OD matrices obtain through the processing of smart card data with a large scale OD survey. The application is done using data from Lyon public transit networks. The research shows that both methods result in similar OD matrices which validate the methodology used to process the smart card data. Yet, some small discrepancies were observed notably, the survey seems to take better account of small trips. Those findings are interesting for public transport operators that plan to replace traditional OD survey by more advanced data processing technics.

Extended abstract:

Origin-Destination (OD) matrices are one of the key elements for the analysis and planning of public transportation networks (Munizaga & Palma, 2012). For decades, transportation researcher have used data of active solicitation such as survey to construct those matrices but the rapid rise of intelligent transport system and recent advances in data collection and data processing are changing this portray (Chen et al., 2016). In this era, AFC (automatic fare collection) data is one of the most promising passively collected data that can be used to build origin-destination matrices (Pelletier et al., 2011). To reach this goal, several researchers have proposed methods and algorithms to estimate the alighting point of a trip in a system where users only validate when boarding (Li at al., 2018) but also to identify transfers and detect activity (Nassir et al., 2015). The data resulting from those two steps is then transformed to derive OD matrix that can be computed at the desired disaggregation level and could be used to replace traditional survey (Munizaga & Palma, 2012). However, there are many aspects that may affect the quality of the resulting matrix such as violation of the alighting inference hypothesis, fare evasion and fare non-interaction, ratio of passenger journeys using smart card data, self-selection bias among those who use smart card data or wrong itinerary choice in subway networks assignment (Hickman, 2016; Munizaga & Palma, 2012; Munizaga et al., 2014; Zhao et al. 2007). Those aspects need to be considered and the resulting matrices need therefore to be validated in an exogenous fashion. To do this some authors have used travel diary survey and count survey (Barry et al., 2002), manually-surveyed bus trips (Wang et al., 2011), data from automatic passenger counting (APC) (Nassir et al., 2011) or even data from volunteers that were asked to validate the results of the model (Munizaga et al., 2014). Those exogenous validations are interesting but they remain partial because of the lack of data. In Lyon, the public transport operator has been for a long time collecting large scale OD survey for the whole network. Given this opportunity, the aim of this paper is to compare the matrices resulting from AFC data with a large scale OD survey. The application is done using one week of AFC data from the transit network of Lyon (between the 13-03-2017 and 19-03-2017). The paper is organized as followed. First, we detailed the case study and the algorithm that are used to transform the raw AFC data into an enrich dataset that allows the creation of OD matrices. Then, we make a series of comparisons with the OD survey to characterize the convergence and the divergence between the two data sources. Finally, we discuss the results of those comparisons with a special emphasis on how we could explain the discrepancy between the two data sources.

TCL ("Transport en Commun Lyonnais") is the commercial name of the public transport network of Lyon. This network is currently run by a private operator under the supervision of the public transport authority of the Lyon metropolitan area (Sytral). The network consists of 4 lines of metro, 2 lines of funicular, 5 lines of tramway and more than 100 regular lines of bus. The current fare transaction system of TCL was implemented in 2002. Smart card and magnetic paper ticket can be used but only smart card can be uniquely identified. Whatever the fare support, passengers are required to validate every time they board a vehicle except in the metro and funicular network where the validation is only needed when entering the system (not tap-in is needed for connections between underground lines). To convert the raw AFC transactions from smart card into an enriched dataset the following steps were implemented:

- Imputation of missing information by merging with AVL (automatic vehicle location);
- Deduplication procedure using heuristics to ensure the integrity of the data;
- Differentiation of transfer transactions from the beginning of new trips using a temporal criteria of 60 minutes between boarding transactions time and a binary criteria comparing the current line (or station) and the next line (or station) (Devillaine et al. 2012);
- Inference of alighting stop or destination station using the trip chaining method (Trépanier et al. 2007; Li at al., 2018) with a maximum walking distance of 600 m and considering for the last validation of the day a return to the closest stop of the first validation of the day.

At the end of this process, an alighting stop or station was estimated for around 80% of the smart card transactions and it was possible to identify the first and the last validation of each trip. This results in approximately 3 million OD pair for the whole week.

In Lyon, all lines are manually surveyed every five years to obtain data regarding the traveller's current trips: origin-destination of the trip leg on the surveyed line, origin-destination of the full trip, socio-demographic, trip purposes and public transit access modality. For bus lines, this survey is done on board and is supposed to be exhaustive. For tramways and metro lines, this survey is done at station and a random sample of approximately 35% of passengers for tramways and 25% of passengers for metro are surveyed. The results are then scaled according to half hour count by stop or station derived from APC (automatic passenger counting) data. Each line is survey for three types of day weekdays, Saturday and Sunday. In order to obtain data for the complete networks, the OD survey from all lines collected between 2012 and 2016 was compiled into one single database that should reflect an average weekday, an average Saturday and an average Sunday and that can be compared to the enrich AFC dataset.

One essential component of transit usage is that the users may need to connect between lines to reach their destination. The first comparison was, therefore, to verify that the number of legs by trips was similar in both data sources. The results show that for this indicators the two data sources are close but the percentage of trips with only one leg is a bit more important in the OD survey dataset (73% vs 68%) which could indicate shorter trips and/or less connections. This difference was further investigated by calculating the distribution of trips distance between the two data sources. The mean Euclidean trip distance for the weekdays was 3293 m in the OD survey compare to 3557 m in the AFC enrich dataset. After that, each data sources was transformed into three OD matrices for each type of day at the communal level. The resulting number of trips was different between each data sources and it was decided to normalize all matrices in order to compare their structure. The results indicate that the two data sources are quite coherent at this level of aggregation. For weekdays, the percentage of pair with GEH inferior to 5 was 83.9% and the R^2 was equal to 0.97. For Saturday and Sunday, the convergence between the data sources remain strong but still a bit weaker than for weekdays as indicated by a decrease in R^2 and in the percentage of GEH inferior to 5. Finally, we compared the total originated trips per commune, the total attracted trips per commune and the total intra communal trips. We observed similar trends in those dimensions however for some communes the AFC data clearly underestimate intra communal trips.

In this work, we report a quantitative validation of the OD matrices obtained with AFC data using a large scale survey. Our results suggest that the matrices estimated using AFC data are consistent with those from the survey but they are some discrepancies that need further investigations. Especially, it seems that the OD obtained with AFC tend to underestimate small trips such as intra communal trips. It is also less accurate on Saturday and Sunday when the use of paper ticket is more important. Those findings could partially be explained by three reasons. First people may not systematically validate when makings short trip. Second tickets users and fare evaders may have different travel characteristics from those who use smart card. Third, during the weekends they may be more violation of trip chaining assumption (such as using other modes of transport) and also more single trips where the alighting point could not be estimated. Based on those observations, we identify the following lines of research: (1) improve the inference methods by using a multi-day approach to improve the percentage of boarding with alighting especially for single trips users (Trépanier et al. 2007), (2) better scaling of the matrices using transaction of tickets users and APC.

References:

Barry, J. J., Newhouser, R., Rahbee, A., & Sayeda, S. (2002). Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record*, *1817*(1), 183-187.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies*, *68*, 285-299.

Devillaine, F., Munizaga, M., & Trépanier, M. (2012). Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, (2276), 48-55.

Hickman, M. (2017). Transit origin-destination estimation. In Kurauchi, F., & Schmöcker, J. D. (2017). *Public Transport Planning with Smart Card Data*. CRC Press.

Li, T., Sun, D., Jing, P., & Yang, K. (2018). Smart card data mining of public transport destination: A literature review. *Information, 9*(1), 18.

Munizaga, M. A., & 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.

Munizaga, M., Devillaine, F., Navarrete, C., & Silva, D. (2014). Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies*, *44*, 70-79.

Nassir, N., Hickman, M., & Ma, Z. L. (2015). Activity detection and transfer identification for public transit fare card data. *Transportation*, *42*(4), 683-705.

Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*(4), 557-568.

Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, *11*(1), 1-14.

Wang, W., Attanucci, J. P., & Wilson, N. H. (2011). Bus passenger origin-destination estimation and related analyses using automated data collection systems. *Journal of Public Transportation*, *14*(4), 7.