Identifying Choice Sets for Public Transport Route Choice Models using Smart-Card data: Generated vs. Empirical Sets

Georges Sfeir[1], Filipe Rodrigues[2], Ravi Seshadri[3], Carlos Lima Azevedo[4]
[1]Technical University of Denmark, Kongens Lyngby, Denmark, geosa@dtu.dk
[2] Technical University of Denmark, Kongens Lyngby, Denmark, rodr@dtu.dk
[3] Technical University of Denmark, Kongens Lyngby, Denmark, ravse@dtu.dk
[4] Technical University of Denmark, Kongens Lyngby, Denmark, climaz@dtu.dk

Keywords: Smart Card data; Choice Set; Route Choice Model; Public Transport

Extended Abstract (2120 words + 2 Figures)

1. Introduction

Public transport route choice models are fundamental components in many transport applications, as they model and predict individuals' route choice behavior and therefore help in assessing and improving public transport network design and performance. A route choice model consists of two main components: 1) choice set generation, which tries to enumerate all possible alternatives between origin and destination pairs; and 2) choice modeling of the chosen alternative from the generated choice set. This study focuses on the first component, as choice sets play a crucial role in understanding travel decision-making behavior. Several studies have shown that the composition and size of choice sets have a significant impact on both choice model estimation and demand prediction (Bovy, 2009; Swait & Ben-akiva, 1987). Inaccurate choice sets could result in misspecification of choice models and introduce biases to forecasted demand (Bovy, 2009; Ortuzar & Willumsen, 2001). Studies on public transport route choice modelling have mainly relied on conventional choice set generation approaches used in road network applications, with some modifications to account for the differences between road and public transport networks (Tan, 2016). Namely, conventional approaches/algorithms such as *k*-shortest path (van der Zijpp & Catalano, 2005), multi-objective path, simulation (Bekhor et al., 2006), branch and bound (Prato & Bekhor, 2006), labeling (Ben-Akiva et al., 1984), link elimination (Azevedo et al., 1993), and/or doubly stochastic (Nielsen, 2000) have been applied in several studies on public transport route choice models to generate exhaustive choice sets (e.g., Abdelghany & Mahmassani, 1999; Anderson et al., 2017; Benjamins et al., 2001; Friedrich et al., 2001; Tan et al., 2015). However, choice set generation is a complex and challenging task in dense urban public transport networks, due to the large number of possible route/path alternatives in such

networks. Enumerating all possible alternatives becomes challenging and impractical. In addition, it is unlikely that all enumerated alternatives are in reality considered by passengers (Gentile & Noekel, 2016). A choice set generation approach should ensure high coverage by generating enough paths to cover passengers' choices but must also ensure high precision by including only paths that are relevant (Marra & Corman, 2020). In addition, the choice set size and quality of the generated alternatives may significantly affect parameter estimates (Frejinger et al., 2009; Zimmermann & Frejinger, 2020). However, defining relevant paths is not an objective task and cannot be easily cross-checked against actual passengers' behavior, which complicates the evaluation of choice set quality. More recently, researchers have relied on observed Smart Card (SC) data, collected by Automated Fare Collection (AFC) systems, to generate choice sets (e.g., Arriagada et al., 2022; Lee & Sohn, 2015; Zhang et al., 2018). The high implementation rate of AFC systems in many countries enables them to cover nearly the entire population of travelers, resulting in substantial volumes of travel data over long periods of time (Bagchi & White, 2005). By using SC data, the observed routes are assumed to form the choice set of the corresponding origin-destination pairs. It is assumed that considering SC data over long periods of time should cover all relevant paths that are considered by passengers.

This study will generate choice sets for a large multimodal public transport network using both conventional approaches and smart card data. It will evaluate and compare the two generated choice sets based on computational performance, coverage tests, and composition tests (size of choice set, diversity of alternatives, variations of path attributes etc.). In addition, route choice models will be developed using the conventional and observed choice sets and compared on the basis of statistical goodness-of-fit measures, interpretation of parameter estimates, and out-of-sample generalization performance.

2. Choice sets generation

2.1. Case Study

This study focuses on the multimodal public transport network (buses, trains, and metros) in the East Great Belt area of Denmark which includes Zealand, the largest island in Denmark and home to its capital Copenhagen. The Danish Rejsekort (travel card in English) is the nationwide smart card system for traveling by public transport in Denmark. Under this system, passengers must tap-in at their origins and transfer locations and tap-out at their destinations. The Rejsekort system covers all public transport modes (buses, trains, and metros), transport operators, and travel zones in Denmark (Rejsekort, 2023). Each Rejsekort transaction stores information on the type of transaction (tap-in, transfer, or tap-out), time and location of the transaction, type of the card, and fake card ID (Rejsekort IDs are pseudo-anonymized for privacy concerns).

2.2. Conventional choice set

The General Transit Feed Specification (GTFS) data that includes stops, service lines, service line frequencies, and travel time information, in addition to the road network from Open Street Map (OMP) were used to build the network and compute path attributes such as in-vehicle travel time, waiting time, number of transfers, walking time and path size (to account for path overlaps). A public transport graph was built containing a set of vertices and a set of edges connecting pairs of those vertices. The vertex set consists of bus stops, train and metro stations, and nodes form the road while the edge set connects the vertices with 4 different edge types: bus, train, metro, and walk. Note that an edge connecting a pair of vertices accounts for all common bus lines serving the same pair. The network consists of 11,419 bus stops, 1,425 bus lines, 244 train and metro stations, 46 train and metro lines, 124,368 nodes (from the road network), and 784,603 road segments. To sum up, the graph consists of 136,031 vertices and 3,513,457 edges. In-vehicle travel time was computed as the average scheduled travel time among all service lines on a specific route segment. A constant walking speed of 4 km/h was considered for walking time calculations. Waiting time was calculated based on the overall frequencies of common service lines on a specific route segment as follows:

$$WT_r = \frac{1}{2\sum_{i=1}^{I} f_i} \quad (1)$$

Where WT_r is the expected waiting time on route segment r, f_i the frequency of line i, and I the total number of service lines on route segment r.

Next, one day of smart card data (19 September 2017) was selected and all observed stop-to-stop (origin-destination - OD) trips were extracted. The one-day Rejsekort dataset contains, after several cleaning steps, 88,700 unique OD pairs. Then, the conventional choice set was created by generating for each OD pair a set of path alternatives using a combination of four choice set generation algorithms/approaches: *k*-shortest path, link elimination, labeling, and simulation. In total, paths for 86,128 unique OD pairs were generated for a coverage rate of 98.88%. The high coverage rate was attained by correcting network errors and using a combination of algorithms/approaches. However, due to the large size of the network, building the network and correcting network errors proved to be a costly and time-consuming task that spanned several months. In addition, the choice set generation required two weeks of computation on a Linux system equipped with CPU @ 2.6 GHz and 196GB of RAM. The generated choice set has only 1.21% of OD pairs with one path/alternative and an average of 7.79 alternatives per OD.

2.3. Observed choice set

For the observed path set, the same Rejsekort day (19 September 2017) was first selected, and paths were generated for each OD pair by considering all the observed trips that start and end within a radius of 400 meters of the origin and destination. A sensitivity analysis will be

performed to define the sensitivity and consistency of the results with respect to the radius. Note that travel time for each path was computed as the average travel time among all trips observed on that path. Number of transfers was also extracted for each path as passengers are required to tap-in at transfers. In addition, some heuristics were applied to check for hidden transfers where people forgot to tap-in. For instance, if a trip consists of two transactions, a tap-in and a tap-out, that happened at two different buses, then it is assumed that one transfer at least occurred on that path.

However, the choice set showed that 54.38% of the OD pairs have only one path/alternative with an average of 2.17 alternatives per OD pair. Therefore, more Rejsekort weekdays were considered to increase the number of paths per OD and generate an exhaustive choice set. By increasing the number of weekdays from 1 to 20, the percentage of OD pairs with only one alternative drops from around 55% to around 14% (Figure 1) while the average number of alternatives per OD increases from 2.17 to 8.59 (Figure 2). A plateau effect could be seen in Figure 1. However, more weekdays will be added to the choice set generation process and the two measures from Figure 1 and 2 will be checked for convergence. In addition, waiting and walking time will be added to the path set. In terms of computational performance, it took less than 5 minutes to generate the choice set with 20 days of data on a Linux system equipped with CPU @ 3.8 GHz and 128GB of RAM.

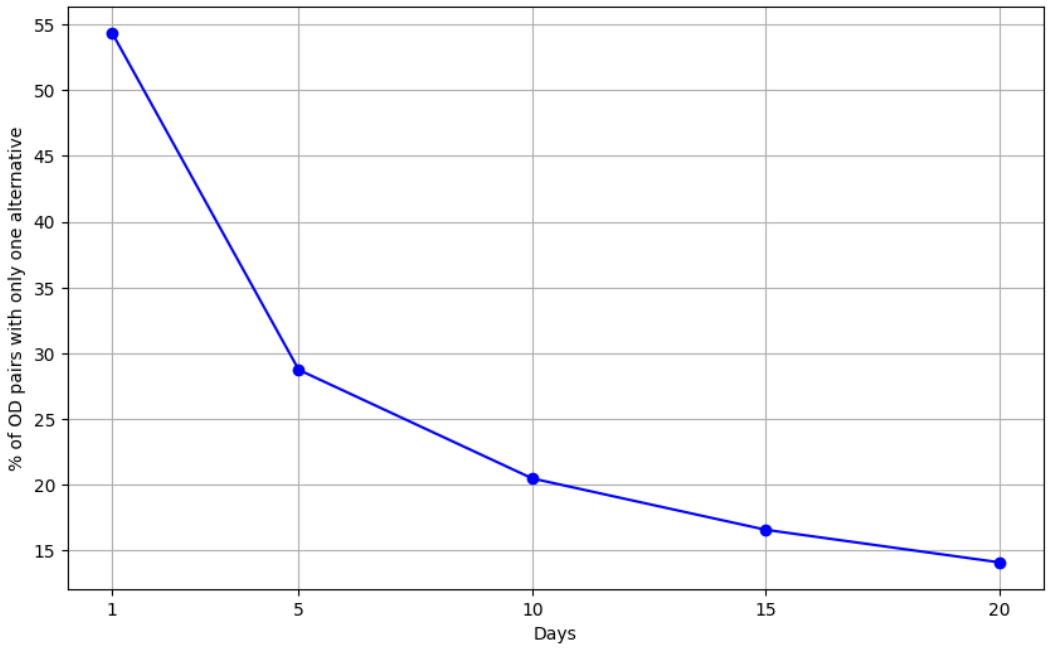

Figure 1: Percentage of OD pairs with one alternative across days

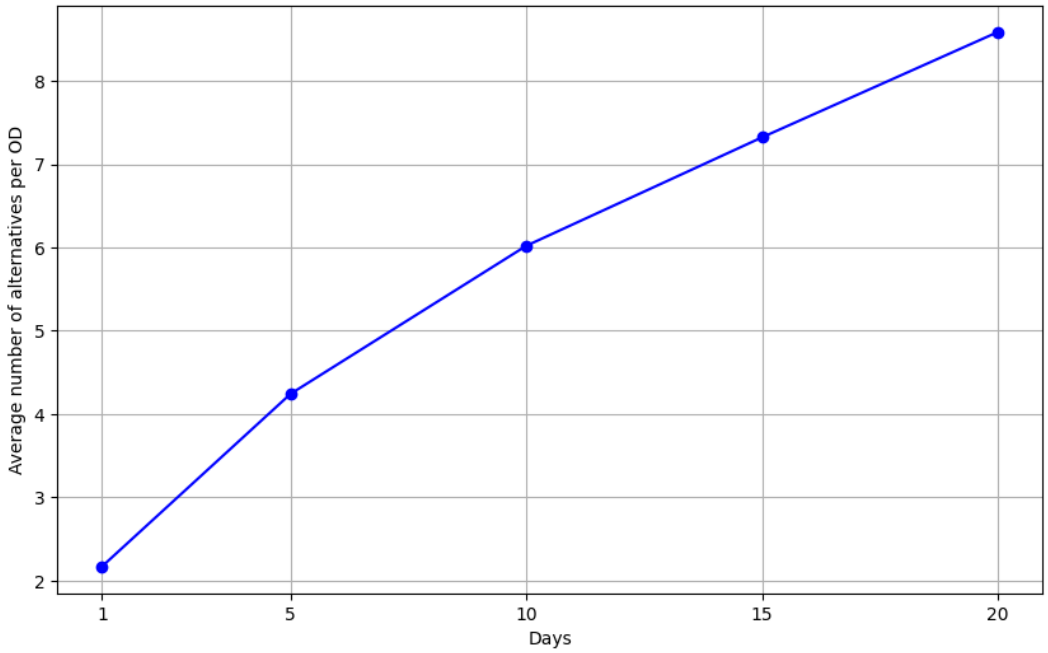

Figure 2: Average number of alternatives peer OD across days

After finalizing the observed choice set based on smart card data as previously mentioned, the observed and conventional choice sets will be compared and evaluated. The evaluation will consider, in addition to the computational performance and the measures from Figure 1 and 2, coverage tests and composition tests such as size of choice set, diversity of alternatives, variations of path attributes, etc. Such measures will help in assessing the quality of the generated choice sets and their suitability for later route choice model estimation. A choice set that has fewer alternatives than the observed ones could introduce biases in the parameter estimates of the route choice model and lead to false predictions/forecasts. Conversely, an excessively large number of alternatives would lead to computational inefficiencies and model estimation challenges (Tan, 2016).

3. Route Choice Modelling

Stop-to-stop multimodal route choice models will be developed using both the conventional and observed generated choice sets to evaluate their impact on the model's goodness-of-fit, parameter estimates and potential biases, in addition to out-of-sample prediction accuracy. Path-size mixed logit models will be developed to account for correlation among alternatives due to path overlapping (Hoogendoorn-Lanser et al., 2005) and heterogeneity across passengers.

In public transport networks, path overlapping is not solely limited to overlapping of alternatives along road segments, but it also includes overlapping of boarding stations (Tan, 2016). At each boarding station, passengers have the flexibility to either continue along their

current path or change to another one. Therefore, path size factors accounting for roads and boarding stations overlaps will be added to the model.
The utility of choosing alternative i from a generated choice set C_n in choice situation n is expressed as follows:

$$U_{in} = \beta'_X X_{in} + \beta_{PS} PS_{in} + \beta_{PS\_node} PS_{in}^{node} + \varepsilon_{in} \quad (2)$$

Where X_{ni} is a vector of attributes for path i, β_X is a vector of corresponding fixed and random coefficients, PS_{ni} and PS_{ni}^{node} are path size factors with β_{PS} and β_{PS_node} their corresponding coefficients, and ε_{ni} is a random disturbance term that is independently and identically distributed (iid) Extreme Value Type I over decision-makers and alternatives.
PS_{in} is a path-size factor that accounts for correlation due to overlapped road segments along path alternatives and is proportional to travel time as follows:

$$PS_{in} = \sum_{r \epsilon \Gamma_i} \left(\frac{t_r}{T_i}\right) ln\left(\frac{1}{\sum_{j \epsilon C_n} \delta_{rj}}\right) \quad (3)$$

Where t_r is the travel time on segment r, T_i is the total travel time of all segments on path i, Γ_i is a set containing all segments along path i, and δ_{rj} is a dummy that is equal to 1 if segment r is part of path i and 0 otherwise.
PS_{ni}^{node} is a path-size factor that accounts for correlation due to overlapped boarding stops/stations as follows:

$$PS_{ni}^{node} = \sum_{s \epsilon S_i} ln\left(\frac{f_{si}}{\sum_{j \epsilon C_n} \gamma_{sj} f_{sj}}\right) \quad (4)$$

Where S_i is the list of all boarding stops in path i excluding origin (initial boarding stop), f_{si} is the boarding frequency over all shared service lines at boarding stop s along path i, and γ_{sj} is a dummy that is equal to 1 if s is a boarding stop in path j and 0 otherwise. Under this formulation, a path containing overlapping boarding stations that are served by more frequent service lines will exhibit a large path-size and as such low negative impact on the overall path utility. On the contrary, a path containing overlapping boarding stations that are served with more frequent service lines will have a smaller path-size leading to more negative impact on the overall path utility.

Finally, this study will compare two approaches for constructing choice sets for public transport route choice models. Through this analysis, we aim to provide insights regarding the optimal circumstances/applications for employing each approach, test their respective modeling strengths and weaknesses, and determine the necessary number of observations (e.g., in terms of days of smart card data) required for generating a choice set from observed smart card data.

Acknowledgement

The research leading to these results has received funding from: 1) the SORTEDMOBILITY project which is supported by the European Commission and funded under the Horizon 2020 ERA-NET Cofund scheme under grant agreement N. 875022; and 2) the Horizon Europe Framework Programme under the Marie Skłodowska-Curie Postdoctoral Fellowship MSCA2021-PF-01 project No 101063801

References

Abdelghany, K., & Mahmassani, H. (1999). Shortest path algorithm for large scale intermodal networks. *INFORMS, Philadelphia, Fall*.

Anderson, M. K., Nielsen, O. A., & Prato, C. G. (2017). Multimodal route choice models of public transport passengers in the Greater Copenhagen Area. *EURO Journal on Transportation and Logistics*, *6*(3), 221–245. https://doi.org/10.1007/s13676-014-0063-3

Arriagada, J., Munizaga, M. A., Guevara, C. A., & Prato, C. (2022). Unveiling route choice strategy heterogeneity from smart card data in a large-scale public transport network. *Transportation Research Part C: Emerging Technologies*, *134*. https://doi.org/10.1016/j.trc.2021.103467

Azevedo, A., Emilia Santos Costa, M. O., Joao Silvestre Madeira, J. E., & Vieira Martins, E. Q. (1993). Theory and Methodology An algorithm for the ranking of shortest paths. *European Journal of Operational Research*, *69*, 97–106.

Bagchi, M., & White, P. R. (2005). The potential of public transport smart card data. *Transport Policy*, *12*(5), 464–474. https://doi.org/10.1016/j.tranpol.2005.06.008

Bekhor, S., Ben-Akiva, M. E., & Ramming, M. S. (2006). Evaluation of choice set generation algorithms for route choice models. *Annals of Operations Research*, *144*(1), 235–247. https://doi.org/10.1007/s10479-006-0009-8

Ben-Akiva, M., Bergman, M., Daly, A. J., & Ramaswamy, R. (1984). Modeling inter-urban route choice behaviour. *The 9th International Symposium on Transportation and Traffic Theory, VNU Press, Utrecht*, 299–300.

Benjamins, M., Lindveld, C., & Van Nes, R. (2001). Multimodal travel choice modelling: a supernetwork approach. *81st TRB Annual Meeting. Washington DC*.

Bovy, P. H. L. (2009). On modelling route choice sets in transportation networks: A synthesis. In *Transport Reviews* (Vol. 29, Issue 1, pp. 43–68). https://doi.org/10.1080/01441640802078673

Description of the Danish Rejsekort, Rejsekort Website, accessed on 09 June 2023. (2023). https://www.rejsekort.dk/.

Frejinger, E., Bierlaire, M., & Ben-Akiva, M. (2009). Sampling of alternatives for route choice modeling. *Transportation Research Part B: Methodological*, *43*(10), 984–994. https://doi.org/10.1016/j.trb.2009.03.001

Friedrich, M., Hofsaess, I., & Wekeck, S. (2001). Timetable-Based Transit Assignment Using Branch and Bound Techniques. *Transportation Research Record: Journal of the Transportation Research Board*, *1752*, 100–107. www.ns.nl

Gentile, G., & Noekel, K. (2016). *Modelling Public Transport Passenger Flows in The Era of Intelligent: COST Action TU1004 (TransITS) Transport Systems* (10th ed., Vol. 1). Cham: Springer International. Print. Springer Tracts on Transportation and Traffic.

Hoogendoorn-Lanser, S., Van Ness, R., & Bovy, P. H. L. (2005). Path Size and Overlap in Multi-modal Transport Networks. *Transportation and Traffic Theory. Flow, Dynamics and Human Interaction. 16th International Symposium on Transportation and Traffic Theory*.

Lee, M., & Sohn, K. (2015). Inferring the route-use patterns of metro passengers based only on travel-time data within a Bayesian framework using a reversible-jump Markov chain Monte Carlo (MCMC) simulation. *Transportation Research Part B: Methodological*, *81*(P1), 1–17. https://doi.org/10.1016/j.trb.2015.08.008

Marra, A. D., & Corman, F. (2020). Determining an efficient and precise choice set for public transport based on tracking data. *Transportation Research Part A: Policy and Practice*, *142*, 168–186. https://doi.org/10.1016/j.tra.2020.10.013

Nielsen, O. A. (2000). A stochastic transit assignment model considering differences in passengers utility functions. *Transportation Research Part B: Methodological*, *34*, 377–402. www.elsevier.com/locate/trb

Ortuzar, J. D., & Willumsen, L. G. (2001). *Modelling Transport, 3rd edn*. (Chichester: John Wiley).

Prato, C. G., & Bekhor, S. (2006). Applying Branch-and-Bound Technique to Route Choice Set Generation. *Transportation Research Record: Journal of the Transportation Research Board*, 19–28.

Tan, R. (2016). *Modeling route choice behaviour in public transport network*.

Tan, R., Adnan, M., Lee, D., & Ben-Akiva, M. (2015). New path size formulation in path size logit for route choice modeling in public transport networks. *Transportation Research Record: Journal of the Transportation Research Board*, 11–18. https://doi.org/10.3141/2538-02

Swait, J., & Ben-akiva, M. (1987). Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B: Methodological*, *21*(2), 91–102.

Van der Zijpp, N. J., & Catalano, S. F. (2005). Path enumeration by finding the constrained K-shortest paths. *Transportation Research Part B: Methodological*, *39*(6), 545–563. https://doi.org/10.1016/j.trb.2004.07.004

Zhang, Y., Yao, E., Zhang, J., & Zheng, K. (2018). Estimating metro passengers' path choices by combining self-reported revealed preference and smart card data. *Transportation Research Part C: Emerging Technologies*, *92*, 76–89. https://doi.org/10.1016/j.trc.2018.04.019

Zimmermann, M., & Frejinger, E. (2020). A tutorial on recursive models for analyzing and predicting path choice behavior. *EURO Journal on Transportation and Logistics*, *9*(2). https://doi.org/10.1016/j.ejtl.2020.100004