Public transport route choice modelling: Identification of bias when using smart card data

Ingvardson, Jesper Bláfoss*[1]; Thorhauge, Mikkel[1]; Nielsen, Otto Anker[1]; Eltved, Morten[2]

[1] Transport Division, DTU Management, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] MOE Tetraplan, Søborg, Denmark

SHORT SUMMARY

Using Automated Fare Collection (AFC) data for public transport analyses has received much research interest recently, including for estimation of passenger preferences through route choice models. However, an important problem persists since AFC data only includes information about the trip within the public transport system, i.e. stop-to-stop. Not knowing the full trip might lead to estimation bias, especially when estimating route choice models using only the chosen stops. This paper highlights this problem by estimating route choice models based on traditional travel survey data and replicated AFC data. In addition, we propose an improved method in which pseudo origin (destination) points in close vicinity of the actually chosen origin (destination) stops are randomly generated, thus allowing pseudo access and egress to be incorporated. The method notably improves parameter estimates of the route choice model compared to estimation assuming AFC stop-to-stop data. Finally, further improvements to the model are presented.

Keywords: AFC data ; Discrete choice modelling ; Estimation bias ; Public transport ; Route choice modelling ; Smart card data.

1. INTRODUCTION

Revealing transport route choice behaviour and preferences among passengers in public transport systems is an important base for evaluating strategies to improve attractiveness of public transport. Such analyses require detailed information on the full trip performed by travellers from their point of origin to their final destination. Traditionally, such analyses have been performed using detailed travel survey data, e.g. (Nielsen et al. 2021; Anderson, Nielsen, and Prato 2017; Berggren et al. 2021; Bovy and Hoogendoorn-Lanser 2005). However, such datasets are costly and limited wrt. sample size. Recently, more focus has therefore been on using automatically collected data from smart card-based automated fare collection systems (AFC). Such data are increasingly used within public transport planning and modelling (Pelletier, Trépanier, and Morency 2011), also for route choice analyses, e.g. (Arriagada et al. 2022; Jánošíkova, Slavík, and Koháni 2014; Nassir, Hickman, and Ma 2019; Raveau et al. 2014; Shelat et al. 2019; Zhao et al. 2017). However, one important drawback of using AFC data for route choice behaviour persists in all these studies, namely the lack of knowledge on the full journey since AFC data only includes trip segments within the public transport system. Hence, the data lacks information on the actual origin and destination, i.e. the access and egress segments of the journey. Not explicitly considering this might introduce bias in the estimation of route choice preferences. An example is when passengers have two options for their first (or last) segment of a trip and can choose to either i) take a bus to the train station, or ii) walk/cycle/drive to the station. If choosing i) then both options (bus and walk/cycle/drive as access to train station) will be available in the choice set for estimating the route choice, whereas if choosing ii) then the choice set will not include option i), since the trip started at the train station according to the AFC data. More specifically, this will introduce biases to the estimates of the in-vehicle times for the various public transport modes where the value of in-vehicle time for bus might be under-estimated.

This study contributes to existing literature within transport route choice in two important aspects. First, by highlighting the potential biases obtained when using AFC data for estimating route choice models in multimodal public transport systems. Second, by proposing a method for estimating route choice models based on AFC data, which reduces estimation bias.

2. METHODOLOGY

For estimating behavioural preferences from transport route choice a traditional two-stage estimation process is applied. This involves i) choice set generation of relevant alternatives to the actual observed choices (CSG), and ii) route choice model estimation.

Choice set generation

The difference between using AFC data and traditional travel survey data lies in the first step. Most previous studies using AFC for route choice analysis have simply generated alternative choice sets based on the revealed stop pairs chosen by the traveller, i.e. points of tap-in and tap-out of public transport, and thus neglected potential access (egress) travel between the chosen stop and the actual origin (destination), as well as possible alternatives at nearby stops in the vicinity of the origin (destination) stop. This study instead suggests to explicitly consider that the point of origin (destination) is not the chosen stop, but rather a point some distance away from the chosen stop. As this point is not known when using AFC data, we propose to simulate random points within a certain distance of the chosen stop, denoted the *sampling distance*, e.g. within a 1,000 meter radius, which is illustrated in Figure 1. These points are used for calculating pseudo access (egress) distances (and travel times) to stops in the choice set (and not only the observed stop). In addition, we suggest to include stops within a certain distance of the chosen stop, denoted the *choice set distance threshold (CSDT)*. This will ensure relevant alternatives in the choice set using other lines than the actually chosen line.

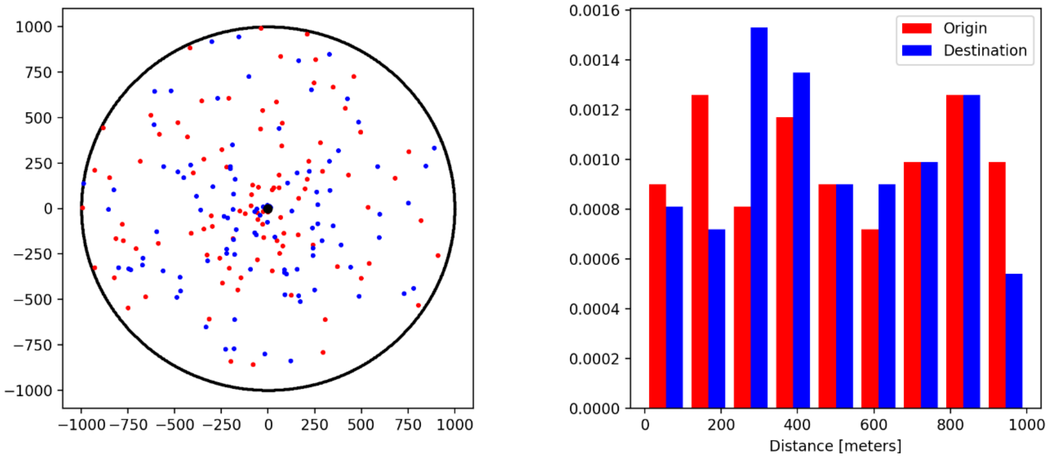

Figure 1: Illustration of the generation of random points (100 draws) in a circle around the actually chosen stop (left) and the corresponding histogram of distances to the center (right)

Model estimation

The utility function for alternative j in the full choice set J_q for individual q is given by:

$$U_{jq} = V_{jq} + \varepsilon_{jq} \tag{1}$$

$$V_{jq} = C_j + \beta_j^Z \cdot Z_{jq} \tag{2}$$

, where U_{jq} is the utility of alternative j for individual q, V_{jq} is the systematic utility, ε_{jq} is a typical i.i.d. Extreme Value (EV) type I error term. C_j is the alternative specific constant for alternative j, Z_{jq} is a vector of level-of-service characteristics of alternative j for individual q, and $ß_j^Z$ its corresponding vector of coefficients. For the level-of-service characteristics we include mode-specific in-vehicle times, walking times and waiting times at transfers as well as a per-transfer penalty term, hidden waiting time, and finally the (pseudo) access and egress times. While the parameter estimates of the (pseudo) access and egress travel times will not be correct, it is hypothesised that the implementation of these variables in the utility function will reduce bias of the estimates of the remaining parameters compared to neglecting these completed.

The model estimation was performed using Monte Carlo simulation in PandasBiogeme using 100 draws (Bierlaire 2020).

3. DATA AND CASE STUDY

The methodology was tested on case study data from the Greater Copenhagen area consisting of a total of 4,810 revealed preference multimodal public transport trips from the Danish travel survey (Christiansen and Baescu 2021). This consisted of 2,553 commuting trips and 2,257 leisure trips. This data contains information about the true origin and destination for each trip, thus the access and egress can be derived precisely. The data was also used in previous studies, which this study builds upon (Anderson, Nielsen, and Prato 2017; Nielsen et al. 2021).

For each observed route choice a number of alternative routes in the multimodal public transport network was generated, which i) used either the chosen stop or a stop in close proximity of the true origin (destination), and ii) at or close to the chosen departure time. The choice set generation was done using full information. For more details we refer to (Rasmussen et al. 2016).

The comparison to estimation based on full knowledge of the entire trip was done by using the full choice set generated in Rasmussen et al. (2016), and then incorporating the *choice set distance threshold (CSDT)*. In the simplest estimation, which was to replicate raw AFC data, we set CSDT = 0 meters, corresponding to only allowing routes between the actually chosen stops. Subsequently, multiple model estimations were done, i.e. CSDT = [100, 250, 500, 1000, 2 500, 5 000, 10 000 and 100 000 meters]. The latter implies using the full choice set. The number of route alternatives in the choice set for each of the model estimations are shown in Figure 2.

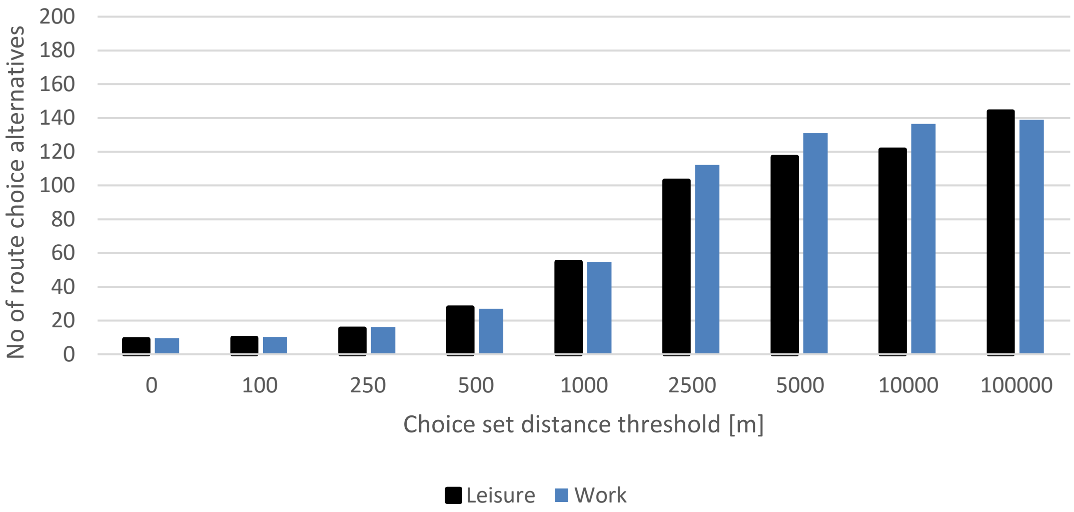

Figure 2: Average number of route alternatives in the choice set for each of the model estimations

When restricting the choice set to only include route alternatives between the actually chosen stops only few relevant routes are included in the model estimation. This might indeed introduce bias to the model estimations since important route alternatives are excluded. For the sample data most relevant alternatives are within 2 500 meters of the chosen stops, and only few are further than 5 000 meters from the actually chosen stops. Such long distances were surprising to observe in the data, but were seen for trips with car as access (egress) to public transport, and especially for leisure trips, such as bringing (picking up) other people.

4. RESULTS

The model estimation results for work trips are shown in Table 1 whereas Table 2 reports rates of substitution. This include the model estimations based on the full choice set, i.e. those when including knowledge on the full door-to-door travel rather than only tap-in to tap-out (*Full knowledge*), including and excluding access/egress parameters, and those estimated based on selected values of the *choice set distance thresholds* (*CSDT*). Figure 3 visualises the accuracy of the parameter estimates for *all* model estimations, i.e. the ratio between the parameter estimate for the given model estimation and the true parameter estimate (from the full choice model). Note here that the ratios representing full data are always equal to one, by definition. From Figure 3 we note that model estimation results stabilise for CSDT at or above 1,000 meters. This is probably due to relevant alternatives most often are within 1,000 meters of the actually chosen stop. Hence, Tables 1 and 2 reports results for the models using CSDT of 250 and 1,000 meters (RoS also reported for 10,000 meters) whereas results for remaining CSDT and leisure are left out due to space restrictions.

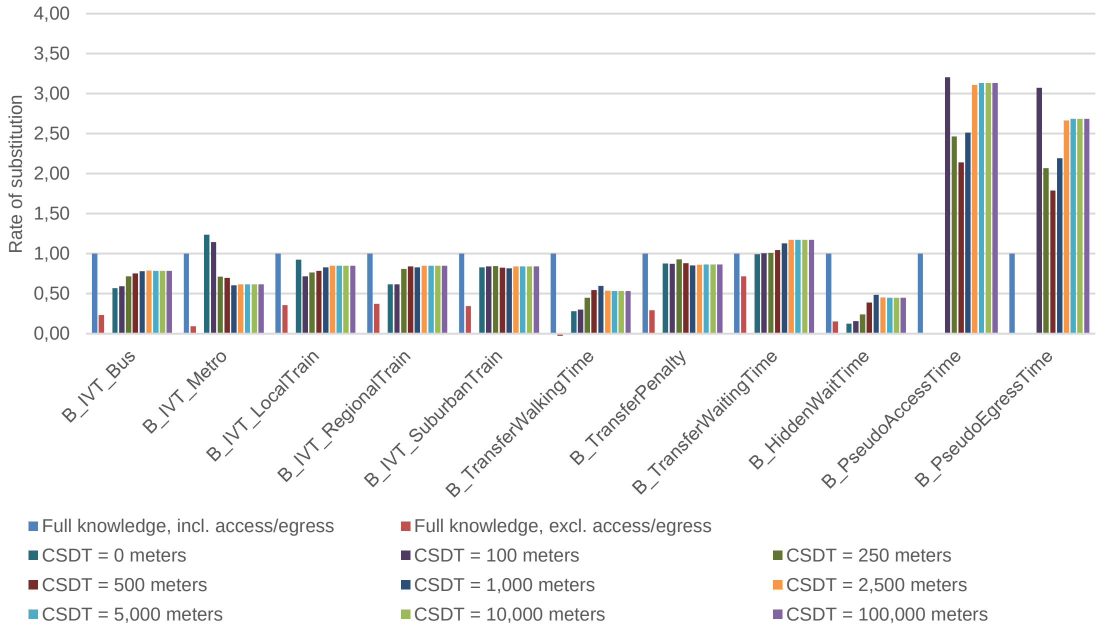

Figure 3: Rates of parameter estimates for the route choice models estimated using various restricted choice sets

The results show that all parameter estimates were highly biased when estimating the model using full knowledge, but excluding access/egress from the model formulation, cf. Figure 3, which is not surprising. When treating the data similar to AFC data and without including stops in vicinity of the chosen stops (CSDT = 0) the model estimates are still highly biased leading to highly biased rates of substitution between level-of-service characteristics.

When adding parameters for (pseudo) access and egress, based on the random origin and destination points, the parameter estimates are still notably off compared to the true estimates. However, accuracy improves by including stops in close proximity, i.e. when increasing CSDT, and becomes stable when including stops within 500-1,000 metes, cf. Figure 3. Further increasing CSDT does not change model estimations notably, probably due to stops further away not being relevant alternatives for the travellers.

Thus, the results suggest that it is very important to include not only routes between the observed stops, but also routes between stops in the vicinity of the chosen stops when estimating route choice models based on AFC data. Even if parameter estimates for the pseudo access/egress coefficients are off, it is important to include them explicitly to improve accuracy of the remaining parameters. However, even though the biases are reduced for in-vehicle times, there are still very large biases related to transfer walking time and hidden waiting time.

Table 1: Estimated parameters for work trips

	Full knowledge		CSDT = 0 meters		CSDT = 250 meters		CSDT = 1,000 meters	
Parameters	**Coef.**	**Rob t-test**	**Coef.**	**Rob t-test**	**Coef.**	**Rob t-test**	**Coef.**	**Rob t-test**
In-vehicle time								
Bus (min.)	-0.313	-20.80	-0.178	-11.4	-0.225	-17.6	-0.244	-23.6
Local train (min.)	-0.274	-9.49	-0.172	-5.27	-0.099	-5.78	-0.084	-6.19
Metro (min.)	-0.139	-6.84	-0.254	-2.62	-0.210	-9.84	-0.227	-11.3
Reg. train (min.)	-0.281	-12.72	-0.173	-7.64	-0.227	-11.7	-0.233	-14.0
S-train (min.)	-0.234	-15.93	-0.194	-12.2	-0.198	-15.3	-0.191	-17.3
Transfer attributes								
Transfer penalty	-2.480	-18.75	-2.18	-17.5	-2.30	-20.3	-2.12	-22.8
Waiting time (min.)	-0.048	-12.69	-0.048	-10.0	-0.049	-10.4	-0.054	-12.7
Walking time (min.)	-0.217	-8.25	-0.061	-2.07	-0.097	-3.81	-0.13	-6.13
Other components								
Access (min.)	-0.488	-18.14	-	-	-1.03	-8.64	-1.05	-26.7
Egress (min.)	-0.418	-17.53	-	-	-1.01	-7.96	-1.07	-25.3
Hidden waiting time (min.)	-0.120	-8.48	-0.015	-1.15	-0.029	-2.04	-0.058	-3.78
Number of est. parameters		11		9		11		11
Number of observations		2,553		2,553		2,553		2,553
Null log-likelihood		-12,589		-4,508		-5,722		-9,235
Final log-likelihood		-2,993		1,703		-1,900		-3,097
Adjusted rho-square		0.761		0.620		0.666		0.663

Table 2: Rate of substitution for estimated parameters for work trips

	Rate of Substitution				
Parameters	**Full knowledge**	**CSDT = 0**	**CSDT = 250**	**CSDT = 1,000**	**CSDT = 10,000 m**
In-vehicle time					
Bus (min.)	1.00	1.00	1.00	1.00	1.00
Local train (min.)	0.44	0.97	0.44	0.35	0.35
Metro (min.)	0.88	1.43	0.93	0.93	0.95
Reg. train (min.)	0.90	0.97	1.01	0.95	0.97
S-train (min.)	0.75	1.09	0.88	0.78	0.80
Transfer attributes					
Transfer penalty	7.92	12.25	10.22	8.69	8.74
Waiting time (min.)	0.15	0.27	0.22	0.22	0.23
Walking time (min.)	0.69	0.34	0.43	0.53	0.47
Other components					
Access (min.)	1.56	-	4.58	4.30	5.33
Egress (min.)	1.34	-	4.49	4.39	5.33
Hidden waiting time (min.)	0.38	0.09	0.13	0.24	0.22

Discussion and future work

While the current methodology notably reduced bias in the parameter estimates as compared to a simple station-to-station model, it can be further improved. In the current method, the choice set distance threshold is based on the distance between *chosen* and *alternative* stops. Another approach is to base CSDT on the distance between the pseudo origin (destination) points and alternative stops. This was not chosen, mainly due to computational considerations, as this requires generation of choice sets for each simulation of the model estimation (100, or preferably 1000 simulations). In addition, the *sampling distance* was kept at 1,000 meters, despite multiple values of CSDT below 1,000 meters. It can be argued that the sampling distance should not exceed CSDT, which will be considered next.

Alternatively, the method can be improved by using a more advanced model, which eliminates the sampling, but rather models the route choice as a conditional probability of the observed stop-to-stop pair, thus incorporating differences in stop choice attractiveness – and in which the choice of origin and destination stops is dependent on service levels between OD-pairs. This elaborated model can be expressed through a nested logit model which include feedback mechanisms from the lower levels (the route choice) to the top level (station choice).

5. CONCLUSIONS

This study has highlighted the problems of estimation bias when estimating route choice models using AFC data. By estimating route choice models based on travel survey data it was possible to replicate model estimations if treating the data as AFC data as well as testing an approach to include alternative stops within certain distances of the chosen stops. The developed framework consisting of random generation of origin and destination points around chosen stops resulted in more accurate model estimations for the level-of-service characteristics, except for access and egress, thus highlighting the importance of including alternative stops when generating choice sets in route choice model estimation using AFC data.

REFERENCES

Anderson, Marie Karen, Otto Anker Nielsen, and Carlo Giacomo Prato. 2017. "Multimodal Route Choice Models of Public Transport Passengers in the Greater Copenhagen Area." *EURO Journal on Transportation and Logistics* 6 (3). Springer Verlag: 221–245. doi:10.1007/s13676-014-0063-3.

Arriagada, Jacqueline, Marcela A. Munizaga, C. Angelo Guevara, and Carlo Prato. 2022. "Unveiling Route Choice Strategy Heterogeneity from Smart Card Data in a Large-Scale Public Transport Network." *Transportation Research Part C: Emerging Technologies* 134 (January): 103467. doi:10.1016/j.trc.2021.103467.

Berggren, Ulrik, Thomas Kjær-Rasmussen, Mikkel Thorhauge, Helena Svensson, and Karin Brundell-Freij. 2021. "Public Transport Path Choice Estimation Based on Trip Data from Dedicated Smartphone App Survey." *Transportmetrica A: Transport Science*. Taylor and Francis Ltd. doi:10.1080/23249935.2021.1973146.

Bierlaire, M. 2020. "A Short Introduction to PandasBiogeme." *Technical Report TRANSP-OR 200605. Transport and Mobility Laboratory, ENAC, EPFL.*

Bovy, Piet H L, and Sascha Hoogendoorn-Lanser. 2005. "Modelling Route Choice Behaviour in Multi-Modal Transport Networks." *Transportation* 32: 341–368.

Christiansen, Hjalmar ;, and Oana Baescu. 2021. *General Rights The Danish National Travel Survey, Catalogue of Variables TU0619v1. Downloaded from Orbit.Dtu.Dk On*.

Jánošíkova, L'udmila, Jiří Slavík, and Michal Koháni. 2014. "Estimation of a Route Choice Model for Urban Public Transport Using Smart Card Data." *Transportation Planning and Technology* 37 (7). Taylor and Francis Ltd.: 638–648. doi:10.1080/03081060.2014.935570.

Nassir, N., Mark Hickman, and Zhen Liang Ma. 2019. "A Strategy-Based Recursive Path Choice Model for Public Transit Smart Card Data." *Transportation Research Part B: Methodological* 126 (August). Elsevier Ltd: 528–548. doi:10.1016/j.trb.2018.01.002.

Nielsen, Otto Anker, Morten Eltved, Marie Karen Anderson, and Carlo Giacomo Prato. 2021. "Relevance of Detailed Transfer Attributes in Large-Scale Multimodal Route Choice Models for Metropolitan Public Transport Passengers." *Transportation Research Part A: Policy and Practice* 147 (May). Elsevier Ltd: 76–92. doi:10.1016/j.tra.2021.02.010.

Pelletier, Marie Pier, Martin Trépanier, and Catherine Morency. 2011. "Smart Card Data Use in Public Transit: A Literature Review." *Transportation Research Part C: Emerging Technologies* 19 (4). Elsevier Ltd: 557–568. doi:10.1016/j.trc.2010.12.003.

Rasmussen, Thomas Kjaer, Marie Karen Anderson, Otto Anker Nielsen, and Carlo Giacomo Prato. 2016. "Timetable-Based Simulation Method for Choice Set Generation in Large-Scale Public Transport Networks." *EJTIR Issue* 16 (3): 467–489.

Raveau, Sebastián, Zhan Guo, Juan Carlos Muñoz, and Nigel H.M. Wilson. 2014. "A Behavioural Comparison of Route Choice on Metro Networks: Time, Transfers, Crowding, Topology and Socio-Demographics." *Transportation Research Part A: Policy and Practice* 66 (1). Elsevier Ltd: 185–195. doi:10.1016/j.tra.2014.05.010.

Shelat, Sanmay, Oded Cats, Niels van Oort, and Hans van Lint. 2019. "Calibrating Route Choice Sets for an Urban Public Transport Network Using Smart Card Data." In *6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*.

Zhao, Juanjuan, Fan Zhang, Lai Tu, Chengzhong Xu, Dayong Shen, Chen Tian, Xiang Yang Li, and Zhengxi Li. 2017. "Estimation of Passenger Route Choice Pattern Using Smart Card Data for Complex Metro Systems." *IEEE Transactions on Intelligent Transportation Systems* 18 (4). Institute of Electrical and Electronics Engineers Inc.: 790–801. doi:10.1109/TITS.2016.2587864.