Public transport trip purpose estimation using automated fare collection data: A comparison of methodological approaches

Ingvardson, Jesper Bláfoss*[1]; Nielsen, Otto Anker[1]; Rich, Jeppe[1]

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

SHORT SUMMARY

Public transport smart card data hold vast amount of information on passenger behaviour. However, no information on trip purpose is recorded, hence limiting its use in practice. While several studies have developed methods for estimating trip purpose, estimation accuracy is still a challenge. This study proposes a two-fold methodology for trip purpose inference, which incorporates i) cluster analysis of trip purposes, and ii) trip purpose estimation. The grouping of similar trip purposes through cluster analysis reduces the complexity of the subsequent trip purpose estimation, hence ensuring a better performance. Several methods are applied for the trip purpose estimation, including discrete choice models with the utility function being specified using Bayesian relevance determination. Preliminary results solely based on temporal characteristics are promising. Future work will include also relevant land use characteristics as well as estimation based on other relevant methods, including Random Forests.

Keywords: AFC data ; Discrete inference ; Public transport ; Smart card data ; Trip purpose estimation.

1. INTRODUCTION

Smart card-based automated fare collection systems (AFC) have been implemented in public transport networks throughout the world during the last decades. This provides basis for collecting much more information about travellers and their preferences on a large-scale using minimum costs. While much travel information is automatically gathered including travel times, route choices and travel frequency (when card id is known) important characteristics often need to be inferred, e.g. actual origin and destination instead of boarding and alighting location (Trépanier, Tranchant, and Chapleau 2007; Munizaga and Palma 2012) and trip purpose (Lee and Hickman 2014; Alsger et al. 2018; Kusakabe and Asakura 2014; Zhu 2020; Kim, Kim, and Kim 2021). These are important characteristics to consider when using this data for estimating and modelling travel behaviour. It is therefore very important to develop efficient methods for post-processing of automated data sources to make the data better suited for behavioural analysis.

Much research has been devoted to inferring trip purpose for automatically collected data. This has contributed to a larger dataset for travel behaviour analysis, hence complementing household travel survey (HTS) data, which traditionally has been the primary source of information used for such analyses. Previously much focus has been on inferring trip purpose for GPS data collected through GPS devices, e.g. (Shen and Stopher 2013), or more recently using smart phones, e.g. (Chen, Jin, and Li 2020). The obvious benefit of this data collection method is that they include all modes of transport, thus complementing HTS data on a large scale. However, an important challenge is the need for efficient post-processing algorithms which are required for identifying

trip segments and transport mode used. The multiple steps required during post-processing before arriving at complete data including trip purpose makes it advantageous to consider other data sources, such as readily available AFC data from public transport. While only covering public transport trips, this data already include the trip chain within the public transport network including boarding and alighting location. This allows for using a readily available dataset for trip purpose inference, thus creating a large-scale dataset for behavioural analysis.

This study contributes to existing literature by proposing a more efficient framework for estimating trip purposes for AFC trips based on household travel survey (HTS) data. More specifically, we propose to group similar trip purposes based on clustering analysis using spatio-temporal trip characteristics rather than using the often large number of trip purposes directly from HTS data. While the grouping of trip purposes that share similar characteristics will reduce the level of detail of the final trip purpose estimation, it has the advantage of reducing the complexity of the subsequent inference model, thus likely resulting in better performance when applied to the AFC data. In addition, the final study will compare the performance of several trip purpose estimation methodologies, including discrete choice models and machine-learning techniques.

2. METHODOLOGY

The proposed model framework consists of three parts, i) clustering of trip purposes, ii) development of an inference model for estimating trip purposes of trips in HTS data, and iii) application of the inference model on AFC data. While it is possible to evaluate the performance of the trip purpose inference model directly as the trip purpose is known for HTS data, this is not possible in the final step using AFC data. Hence, this is proposed done aggregately by comparing the trip purpose splits of the AFC and HTS data.

2.1 Clustering of trip purposes

Most HTS data include a large variety of trip purposes. It will therefore be reasonable to group trip purposes to reduce the complexity of the inference model. The main advantage of this is that the performance of the inference model is expected to improve. This might be at the expense of less details in the trip purpose inference. However, this is suggested as an improvement considering that better prediction accuracy of a simpler trip purpose categorisation is preferred over worse prediction accuracy of a more advanced categorisation. Notably, a more advanced categorisation will result in larger difficulties in separating trip purposes that share similar characteristics. This might not be necessary when using the resulting data for travel behaviour analysis or transport models, which often have fewer trip purpose categories than often reported in HTS data.

The framework proposed in this study is to apply hierarchical clustering analysis (HCA) on the HTS data based on spatio-temporal characteristics of the trips. These are characteristics that are highly linked to various activity types while being simple to implement in a clustering algorithm. Several trip characteristics and indicators calculated for each trip purpose were considered and tested, including the time of travel (e.g. day and hour), time spent on destination (e.g. time between consecutive trips), and land use information at destination (e.g. land use type, number of workplaces and inhabitants at destination). This resulted in the use of the following indicators: i) average trip departure time, ii) average time spent at destination, iii) the share of trips performed on weekdays, and iv) the ratio between workplaces and inhabitants at the destination.

2.2 Trip purpose inference model estimation

We will test three methods for estimating the trip purposes. First, discrete choice models based on an automatic method for specifying the utility functions of the various trip purposes based on Bayesian relevance determination of the variables (including interactions) from the dataset (Rodrigues et al., 2020). This allows for automatic selection of the key spatio-temporal variables in the HTS data, which are correlated with each of the trip purposes. This can be compared directly to specifying the utility functions based on simple backward elimination. Second, we will test a novel approach based on embeddings, which is advantageous in efficiently reducing large sets of categorical/binary variables to a lower dimension vector of continuous numbers while simultaneously considering the individual relationships between the original categorical variables. As the transformation of the original variables to a continuous vector representation is performed with explicit consideration of the goal of inferring trip purpose this method might perform better than previously applied methodologies. Lastly, as several studies found Random Forest models to perform better than alternative approaches (Ermagun, Fan, Wolfson, Adomavicius, & Das, 2017; van Dijk, 2018), we will compare our results to such approach.

2.3 Trip purpose inference model application

Considering that the AFC data is longitudinal it holds more information than is available in the HTS data. This allows for incorporating a final step in the trip inference model in which a set of travel characteristics are identified for each user in the AFC data, e.g. most used stops/stations. This information is used to update the trip purpose estimated by the inference model based on a set of relevant rules.

3. CASE STUDY AND DATA

The methodology is applied on data from the Greater Copenhagen area, which represents a region where public transport has a market share of 10-20% dependent on trip purpose (Christiansen and Skougaard 2015). The main dataset for the trip purpose estimation is the Danish national travel survey, which included 14,714 public transport trips for the period 2006-2019. While the travel survey data include detailed background information on respondents we only use information that can be re-generated for the AFC data, i.e. the spatial and temporal trip data. This includes the trip departure time and time between consecutive trips (time spent at destination/activity), which is taken directly from the HTS data. In addition, we collected land use data consisting of a categorical land use type as well as population and amount of workplaces, all in GIS zone layers. For each public transport stop the corresponding land use type and population and job density could be included. Job density could be used both aggregately and specified for nine different job types.

The AFC data used in this study is the Rejsekort data from the Greater Copenhagen area. This AFC data includes full information on the entire public transport trip, as it is validated at the boarding stop, potential transfer stops and at the alighting stop. This ensures the full route of the passenger within the public transport system. The dataset for 2019 includes more than 120 million trips, which corresponds to approximately 50% of all public transport trips in 2019 (Christiansen and Baescu 2021). However, the sample is biased due to elderly, students, and commuters using public transport on a daily basis can use other types of ticket products at a lower fare. Hence, the sample includes irregular travellers and those that use public transport up to a few times per week (Eltved et al. 2021).

4. RESULTS

The preliminary results are based on temporal trip characteristics and limited spatial characteristics, as detailed spatial characteristics were not available yet. Table 1 shows the results of the HCA based on i) trip departure time, and ii) time between consecutive trips, resulting in six trip purpose categories.

Table 1: Overview of travel survey data used for the trip purpose inference modelling

Trip purpose	No of obs	Cluster	Name	No of obs
Work	2,709	1	W(ork)	2,709
Education	1,268	2	Edu(cation)	1,268
Shopping	1,161	3	Shop(ping)	1,422
Healthcare	261	3		
Pick-up/bring persons	160	4	Vis(its)	1,262
Private visits	1,102	4		
Pickup/bring stuff	93	5		
Other errands	139	5		
Sports	274	5		
Entertainment	883	5	Lei(sure)	1,757
Vacation, excursion	59	5		
Private meeting	152	5		
Other leisure	157	5		
Home	6,296	6	H(ome)	6,296

Subsequently, two trip purpose inference models were estimated based on discrete choice models using i) simple backward elimination, and ii) Bayesian relevance determination. The models were trained on 80% of the sample data (11,833 observations) drawn randomly before testing the inference models on the test dataset (2,881 observations). The results are shown in Tables 2 and 3.

Table 2: Confusion matrix for trip purpose inference based on backward elimination

Real / predicted	W	Edu	Shop	Vis	Leis	H	Sensitivity
Work	394	63	4	9	24	37	531 (74%)
Education	73	107	3	5	20	10	218 (49%)
Shopping	6	5	172	0	64	54	301 (57%)
Visits	14	10	21	13	49	134	241 (5%)
Leisure	15	11	61	11	157	111	366 (43%)
Home	14	4	27	9	40	1,130	1,224 (92%)
Precision	516	200	288	47	354	1,476	2,881 (68%)
	(75%)	(54%)	(60%)	(28%)	(44%)	(77%)	

Table 3: Confusion matrix for trip purpose inference based on Bayesian relevance determination

Real / predicted	W	Edu	Shop	Vis	Leis	H	Sensitivity
Work	384	64	10	5	21	47	531 (72%)
Education	79	107	11	3	7	11	218 (49%)
Shopping	16	8	155	1	49	72	301 (52%)
Visits	17	10	23	24	32	135	241 (10%)
Leisure	37	7	99	15	80	128	366 (22%)
Home	15	5	26	27	38	1,113	1,224 (91%)
Precision	548	201	324	75	227	1,506	2,881 (65%)
	(70%)	(53%)	(48%)	(32%)	(35%)	(74%)	

The results show that promising results considering that no spatial characteristics were include yet. The estimation accuracy of commuting trips are highest at 70-74% accuracy for work- and home-bound trips. The remaining trip purposes are notably lower at 32-43%. The results are compared to related studies in Figure 1.

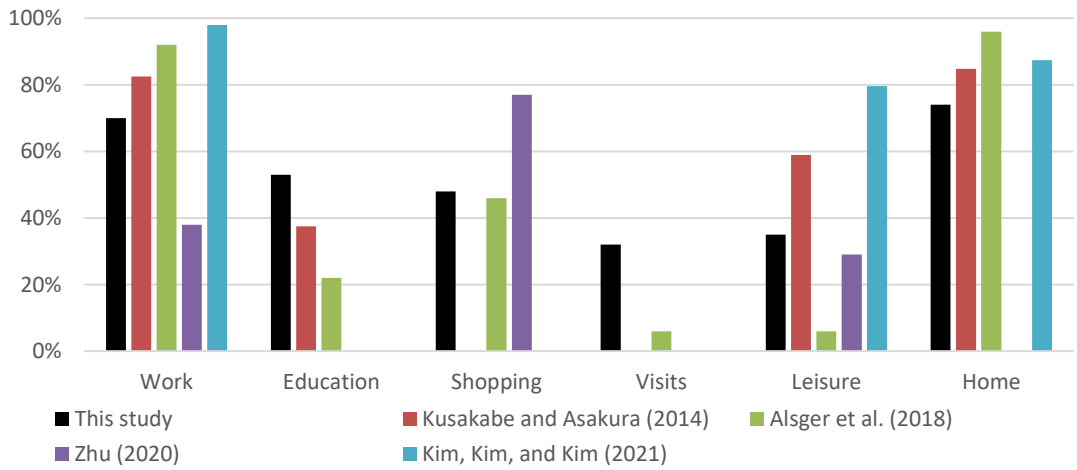

Figure 1: Comparison of results to other related studies

The related studies did not use similar trip purposes, so not all results can be compared consistently. However, the work- and homebound trips are estimated at a lower accuracy than competing studies whereas the remaining four categories are at similar or even better accuracy.

5. DISCUSSION AND FUTURE WORK

As this is very preliminary work, future work will include adding detailed spatial information based on readily available GIS data. This was not possible to include before the deadline of this short paper, but will be included in the final paper. In addition, additional estimations of the trip purpose inference models will be conducted, including using Random Forests and embeddings.

ACKNOWLEDGMENT

The authors would like to thank Independent Research Fund Denmark (DFF) for funding this study as part of the International Postdoc funding program.

REFERENCES

Alsger, Azalden, Ahmad Tavassoli, Mahmoud Mesbah, Luis Ferreira, and Mark Hickman. 2018. "Public Transport Trip Purpose Inference Using Smart Card Fare Data." *Transportation Research Part C: Emerging Technologies* 87 (February). Elsevier Ltd: 123–137. doi:10.1016/j.trc.2017.12.016.

Chen, Yanyan, Zeqian Jin, and Chen Li. 2020. "Trip Purpose Prediction Based on Hidden Markov Model with GPS and Land Use Data." In *2020 IEEE 5th International Conference on Intelligent Transportation Engineering, ICITE 2020*, 55–59. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ICITE50838.2020.9231419.

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

Christiansen, Hjalmar, and Britta Zoëga Skougaard. 2015. "Documentation of the Danish National Travel Survey" 10 (June): 1–18.

Eltved, Morten, Nils Breyer, Jesper Bláfoss Ingvardson, and Otto Anker Nielsen. 2021. "Impacts of Long-Term Service Disruptions on Passenger Travel Behaviour: A Smart Card Analysis from the Greater Copenhagen Area." *Transportation Research Part C: Emerging Technologies* 131 (October). Elsevier Ltd. doi:10.1016/j.trc.2021.103198.

Kim, Eui Jin, Youngseo Kim, and Dong Kyu Kim. 2021. "Interpretable Machine-Learning Models for Estimating Trip Purpose in Smart Card Data." *Proceedings of the Institution of Civil Engineers: Municipal Engineer* 174 (2). ICE Publishing: 108–117. doi:10.1680/jmuen.20.00003.

Kusakabe, Takahiko, and Yasuo Asakura. 2014. "Behavioural Data Mining of Transit Smart Card Data: A Data Fusion Approach." *Transportation Research Part C: Emerging Technologies* 46. Elsevier Ltd: 179–191. doi:10.1016/j.trc.2014.05.012.

Lee, Sang Gu, and Mark Hickman. 2014. "Trip Purpose Inference Using Automated Fare Collection Data." *Public Transport* 6 (1–2). Springer Verlag: 1–20. doi:10.1007/s12469-013-0077-5.

Munizaga, Marcela A., and Carolina Palma. 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. Elsevier Ltd: 9–18. doi:10.1016/j.trc.2012.01.007.

Shen, Li, and Peter R. Stopher. 2013. "A Process for Trip Purpose Imputation from Global Positioning System Data." *Transportation Research Part C: Emerging Technologies* 36. Elsevier Ltd: 261–267. doi:10.1016/j.trc.2013.09.004.

Trépanier, Martin, Nicolas Tranchant, and Robert Chapleau. 2007. "Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System." *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 11 (1): 1–14. doi:10.1080/15472450601122256.

Zhu, Yi. 2020. "Estimating the Activity Types of Transit Travelers Using Smart Card Transaction Data: A Case Study of Singapore." *Transportation* 47 (6). Springer: 2703–2730. doi:10.1007/s11116-018-9881-8.