Public Transport Route Choice Model Based On Feature Importance Derived From Clustering Analysis

Gal Shachar Bekerovich*[1] and Shlomo Bekhor[2]

[1,2]Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology,
Haifa 32000, Israel
[1,2]Emails: galshachar@campus.technion.ac.il, sbekhor@technion.ac.il

Short summary

This paper examines the importance of public transport's unique characteristics in the estimation of public transport route choice models by applying clustering analysis to identify significant patterns in the sampled data. The proposed methodology employs a data-driven approach for generating the choice set and for characterizing the important explanatory variables in the route choice model. The utility specification of the public transport route choice model is derived from the feature importance obtained in the clustering results. The feature selection based on the clustering yields significant explanatory variables in the public transport route choice model.

Keywords: data-driven, machine learning, public transport, route choice.

1 Introduction

The significance of understanding route choice behavior is well-known in transportation. Many studies have been conducted over the years, investigating the problem in a variety of aspects, focusing mainly on car drivers' behavior (Prato, 2009). The specification and estimation of route choice models are not trivial tasks. The choice model itself includes a choice set generation process, which is required to infer the alternatives considered by the individuals. This process is very demanding, due to the large number of possible alternatives (Prato, 2009), and it is more complex in public transportation. This is because of the additional travel characteristics that should be considered in the route choice process, such as the number of transfers, access or egress distances, and transit frequency. Those characteristics have a great influence on the passengers' behavior, as discussed in Anderson (2013). Understanding the preferences of public transport users and the impact of the additional travel characteristics is essential for public transport network and service planning.
Recent studies have demonstrated the capabilities of utilizing data-driven methods to deal with complex problems. For example, Elizalde-Ramírez et al. (2019), proposed a method that aims to capture the travel plan opportunities in the public transport network using artificial intelligence (AI). Formulating public transport characteristics as features for the application of AI and Machine Learning (ML) tools enables further research on route choice behavior, such as route choice model estimation. Arriagada et al. (2022) tried to reveal strategy heterogeneity among public transportation users using smart-card data. Marra and Corman (2020) tried to find an efficient choice set algorithm for multimodal public transportation based on Automatic Vehicle Location data; this data allows for precise mapping of the alternatives but requires additional knowledge of the users' choices, in contrast to GPS trajectory. Tomhave and Khani (2022) discussed the importance of the unique characteristics in public transport route choice, and based their work on on-board (OB) survey data, while Yao and Bekhor (2020) presented a data-driven choice set generation method, in which the choice set was sampled using clusters created from household travel survey data. However, their work was conducted only for private car mode.

Although route choice problems have been extensively researched over the years, data-driven approaches to such models in the context of public transportation have not been fully explored. This study aims to examine the importance and influence of public transport's unique characteristics on the route choice model estimation. Specifically, the paper will use feature importance derived from the application of clustering on household survey data and include the significant features in the public transport route choice model. The paper provides additional insights into the passengers' behavioral preferences that contributed to the route choice model explainability by utilizing the clustering analysis results.

2 Methodology

The suggested methodology employs a data-driven approach for generating the choice set alternatives and for characterizing the important explanatory variables for the route choice model.
This methodology relies on five major steps: data preparation, choice set generation, feature selection, model specification, and model estimation.

Data Preparation

The data used in this paper is based on the Tel-Aviv Metropolitan household travel survey from 2016-2017. A detailed description of the data collection process can be found in Nahmias-Biran et al. (2018). This data contains GPS samples and household information on the individuals in the survey. In addition, the highway and transit networks including all service lines of the Tel-Aviv metropolitan area were used in the process.
The public transport route choice model is derived only for trips conducted in this mode; therefore, only trips made by the passengers that choose public transport as their mode were considered. In this paper out of 3,257 trips identified as public transport trips, 1,898 were processed by manually map-matching the GPS trajectories to the network. From the map-matched trips, 1,007 were found valid for clustering analysis, containing at least 70% overlap with the real observations in the dataset. Finally, 882 trips were used in the final public transport route choice model, which included at least two alternative routes.
Table 1 contains overall statistics from the 1,007 valid observations. The average number of trips per person is equal to 1.8, a reasonable value that indicates that the data set contains both trips in different directions per day. 16.5% of the trips contained transfers and overall we had inter-urban and urban trips in the data set.

Table 1: Observation Statistics

Observation Statistics	
Unique users	558
Trips with transfers	167
Avg. number of trips per person	1.8
Min trip Length	0.5 Km
Max trip Length	66.8 Km
Avg. trip Length	6.4 Km

Choice-set generation

For the 1,898 map-matched routes, the route choice set was generated using the Pathfinder algorithm (Dial, 1967) that was implemented in the commercial software package TransCAD. We considered three different scenarios for the transit penalty settings (walking time weights and transfer weights) in the algorithm, using multiple scenarios for each penalty combination. By running different scenarios, we were able to generate additional routes to the deterministic Pathfinder algorithm. In this way, we can detect diverse alternatives in the choice set, which is essential for covering different characteristics and accounting for the real-world complexity into the process of users' choices.

Feature Selection

For the map-matched trips, we run the K nearest neighbors (Knn) clustering method (Fix & Hodges, 1989). As discussed in Yao and Bekhor (2020), in order to properly perform clustering, there is a need to normalize the route characteristics attributes, used as features in the algorithm.

Another justification for the normalization stems from the variability of the origin-destination (O-D) pairs in the dataset with respect to their different lengths.

Eight different route characteristics attributes were examined, partially described as follows:

Normalized Route Length

The normalized route length is the ratio between the length of route i, L_i^{od} to the maximum route length $\max\left(L^{od}\right)$, in the dataset.

$$\text{Route Length} = \frac{L_i^{od}}{\max\left(L^{od}\right)} \tag{1}$$

Normalized Route Directness

Route directness is the ratio between the route length L_i^{od} to the Euclidean distance between the origin and the destination $\|o-d\|$ for each trip. The lowest possible value for the Route directness is 1. Public transport routes with a relatively high route directness ratio indicate more curvature to other alternatives or even other modes.
The normalized route directness is obtained by dividing each route directness to the maximum value in the dataset.

$$\text{Route Directness} = \frac{L_i^{od}}{\|o-d\|} / \max\left(\frac{L^{od}}{\|o-d\|}\right) \tag{2}$$

Normalized Route Frequency

The route frequency normalized to the highest frequency in the dataset. The frequency value considered is the access frequency at the origin stop.

$$\text{Route Frequency} = \frac{\text{Frequency}_i^{od}}{\max\left(\text{Frequency}^{od}\right)} \tag{3}$$

Additional normalized route characteristics included in the model was Normalized Route number of transfers, Normalized Route transfer distance and Normalized Route access/egress distance.

Additional characteristics that were examined included station capacity at the origin and different combinations of the walking distance attributes. However, these characteristics were not significant in the clustering process and therefore were excluded in the model estimation process.

Model specification and estimation

For simplicity, this paper compares two simple model forms, the Multinomial Logit (MNL) and the Path Size Logit (PSL) models. Model estimation was performed using Pandas Biogeme (Bierlaire, 2003).

3 Results and discussion

Choice Set Generation

Figure 1 shows an example of the choice set generation for a single O-D pair. the left-hand side of the figure corresponds to the pathfinder algorithm with a fixed set of parameters. The right-hand side of the figure corresponds to the pathfinder algorithm with the combination of different parameters set as indicated in the methodology section. The results show that when combining routes obtained from different parameter sets, we enriched a higher number of alternatives, along with the capture of the chosen alternative (e.g., the map-matched route) within the generated choice set.

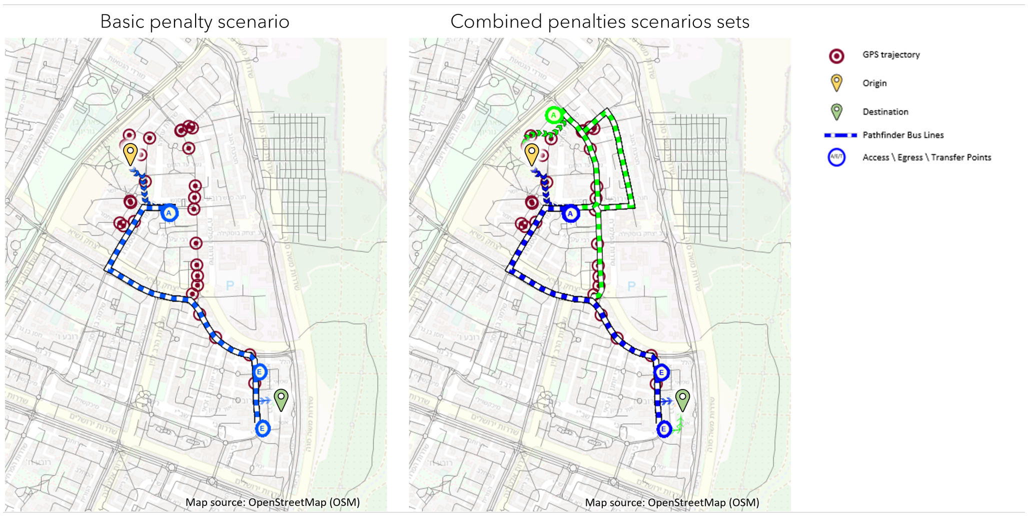
Basic penalty scenario
Combined penalties scenarios sets
GPS trajectory
Origin
Destination
Pathfinder Bus Lines
Access \ Egress \ Transfer Points
Map source: OpenStreetMap (OSM)
Map source: OpenStreetMap (OSM)

Figure 1: Choice set generation example.

The overlap percentage (Bekhor et al., 2006) of one alternative compared to the map-matched route is defined in the following equation:

$$\text{Overlap}_j = \frac{\sum_{a\in\Gamma_j} l_a}{L_i}, \tag{4}$$

Where $\sum_{a\in\Gamma_j} l_a$ is the sum of all common links length between the generated alternative j and the chosen alternative, and L_i is the total length of the chosen alternative.

In the example presented in Figure 1 the overlap percentage of the simulation method is higher than the overlap percentage received from the deterministic choice set generation.
Table 2 shows the overlap rates of the generated choice sets obtained from the combined scenarios results, for threshold values of 70%, 80%, 90% and 100%. The 100% overlap rate means that the real chosen path was fully captured by the generated choice set for the O-D pair. This framework was discussed and implemented on private cars by Bekhor et al. (2006), in which 80% overlap threshold is a sufficient indication of the chosen route for private cars. However, considering the Israeli relatively sparse public transport network, with GPS trajectories from the users' devices alone, we have decided to use a threshold of 70% for the identification of the chosen alternative from the generated choice set, reaching to value of 57% choice sets over this rate. A possible explanation for the low coverage may be related to the network resolution; this paper uses the full highway network, as opposed to previous papers that used planning networks containing only the main network streets.

Table 2: Overlap rate results

Scenario	**Overlap rate for coverage**			
	70%	**80%**	**90%**	**100%**
Basic penalty	49.4%	40.9%	29.0%	11.7%
Transfers sensitivity penalty	49.6%	40.9%	28.1%	11.5%
Basic penalty	50.2%	42.4%	30.7%	12.5%
Combined set from all	**57.3%**	**45.6%**	**32.9%**	**13.3%**

Route Characteristics Attributes Clustering

Using the normalized route characteristics attributes as the clustering features, we run the Knn algorithm with a different number of clusters k, from 2 to 15, aiming to find the optimal value of k that maximizes the Silhouette score (Rousseeuw, 1987) while observing a plateau in the within-cluster sum of squares (WCSS) score (Thorndike, 1953). Figure 2 shows that the selection of k = 4 can be appropriate for our sample.

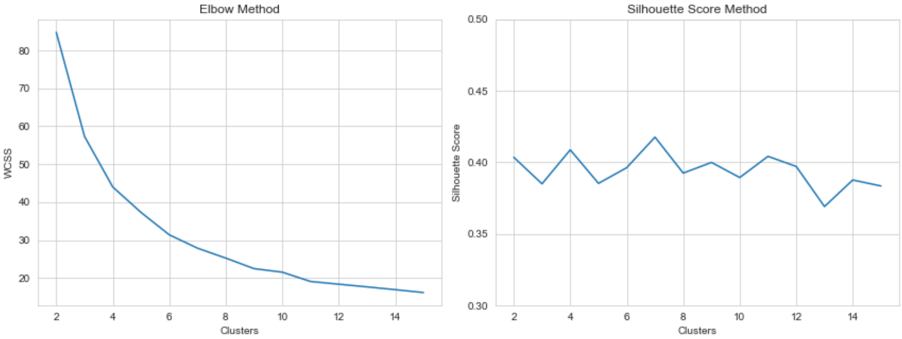

Figure 2: silhouette and WCSS scores over number of clusters.

Table 3 presents general statistics of the normalized route characteristics attributes used as features in the clustering algorithm. The relatively low average value for the number of transfers indicates that the passengers prefer trips without transfers. The maximum number of transfers found in the dataset was 2 transfers, this can be explained by the fact that all observations in this dataset are related to bus trips.

Table 3: Route characteristics attributes – general statistics

Route characteristics	**Min**	**Max**	**Average**	**Standard deviation**
Route length	0.007	1.00	0.095	0.125
Route frequency	0.111	1.00	0.481	0.229
Number of transfers	0.000	1.00	0.149	0.235
Access/Egress walking distance	0.004	1.00	0.068	0.050
Transfer walking distance	0.000	1.00	0.017	0.056
Route directness	0.278	1.00	0.306	0.052

Feature Importance

One of the most important aspects of ML algorithms implementation is the ability to explain the results. Feature importance is an explainability tool for the clustering process, it is used to understand which feature contributes the most to the formation and differentiation of the clusters. By understanding the importance of these features for each cluster, we will gain prior knowledge and use it in the route choice model estimation phase. Figure 3 shows the feature importance results for the 4 clusters created using the Knn algorithm (Fix & Hodges, 1989).

The results of the four graphs in Figure 3 show that the frequency feature was the most important predictor for cluster membership. While, for the rest of the features, the decision rule is less consistent and changes from one cluster to the others. Yet, additional important features for the clustering were the route directness and the number of transfers. Clusters 1 and 2 represent direct trips without transfers, while the difference between them lies in the walking distance of access and egress. This difference can be interpreted as an indication of the passenger's willingness to walk to the station as a function of the trip length. Cluster 3 represents longer trips with transfers and less frequent routes. it contains mainly the inter-urban trips in the dataset. In this aspect, frequency is important because longer trips with transfers tend to be less appealing to the user. Therefore, high frequency in these alternatives can be a decision factor for choosing this category as we can see in cluster 4.

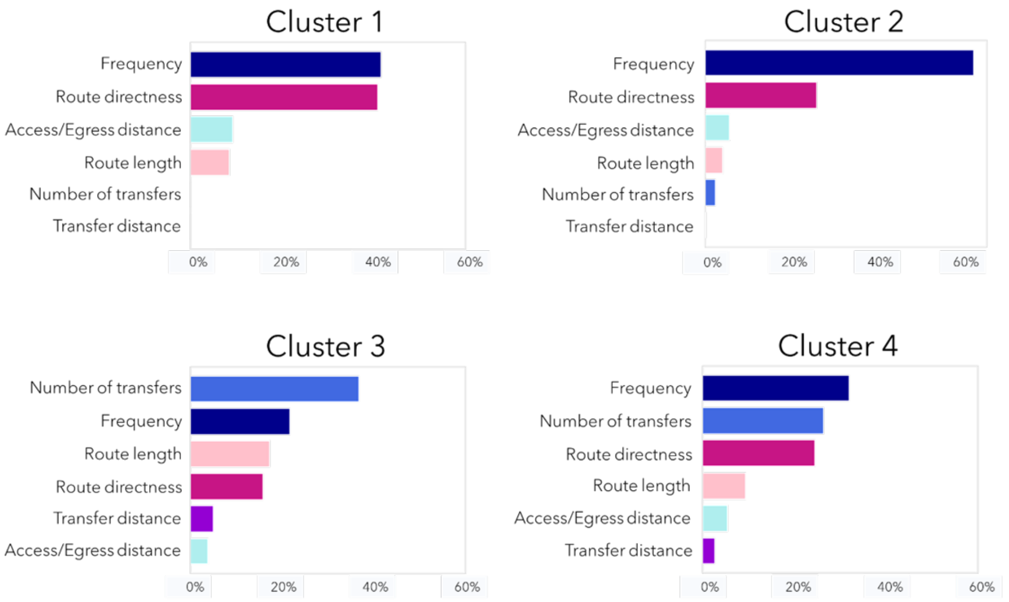

Figure 3: Feature importance for each cluster.

Model Estimation

In this paper, we estimated MNL and PSL public transport route choice models with the features extracted from the clustering analysis as explanatory variables. The estimation results are presented in Table 4 with the coefficient values and the matching p-values for them.

Table 4: model estimation results

	MNL		**PSL**	
	LL(0) = -1155.938 LL(β) = -865.912 Rho-square: 0.217 Rho-square-bar: 0.211		LL(0) = -1155.938 LL(β) = -846.154 Rho-square: 0.235 Rho-square-bar: 0.228	
Coefficient	value	p-value	value	p-value
Route directness	-8.98	0.00	-10.4	0.00
Route length	-4.55	0.01	-3.88	0.03
Route frequency	5.45	0.00	5.46	0.00
Number of transfers	-2.42	0.00	-1.92	0.00
Access Egress walking distance	-9.76	0.00	- 8.41	0.00
Transfer walking distance	-0.87	0.12	-0.78	0.19
PSC			5.66	0.00

The results demonstrated significant p-values for all coefficients except for the transfer walking distance for both models. Our interpretation is that there might not have been enough variability in the dataset concerning this aspect, as discussed earlier regarding the number of transfers characteristic statistics.

The addition of the PSC coefficient in the PSL model enhanced the overall model; Note that this coefficient is positive, which indicates that passengers' prefer routes with common links. This contrasts with results obtained for private car route choice (Bovy et al., 2008). Unlike private car mode, in which disjoint routes are more clearly perceived, in public transport, joint routes can be an indicator of a corridor service, with additional opportunities and more accessibility.

4 Conclusions

This study examines the influence of public transport's travel characteristics on route choice modelling. The approach of the study was to use a comprehensive household travel survey that included route choice data collected from GPS devices. The dataset created included a sufficient number of observations for clustering analysis and model estimation.
To maximize the utilization of the map-matched observations, a high overlap percentage is required. The suggested methodology was able to demonstrate that conventional methods have the potential to cover the observed alternatives. Yet, further research is necessary to reach higher overlap results. The clustering step enabled the identification of important features that were subsequently applied in the model estimation step. The use of normalized explanatory variables in route choice models is not a common issue, and this paper expanded the method proposed by Yao and Bekhor (2020) by applying it on public transport characteristics. The variable normalization is essential for obtaining significant clustering results.
The model estimation results show that the proposed methodology can reach significant parameter estimates respectively to the data. In addition, the feature importance obtained the clustering step provided new insights into the passengers' behavioral patterns, and enriched the route choice model results from the behavioral understanding aspect.
In this paper we estimated simple MNL and PSL choice models. The variables used for estimation were significant and in line with the feature importance results. The proposed methodology can accommodate other complex model forms that may be suitable for the public transport dataset. Further research will investigate additional model forms and include personal explanatory variables from the household survey.

Acknowledgements

This work was supported in part by the Israeli Smart Transportation Research Center.

References

Anderson, M. (2013). *Behavioural models for route choice of passengers in multimodal public transport networks* (Unpublished doctoral dissertation).

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*, 103467.

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*, 235–247.

Bierlaire, M. (2003). Biogeme: A free package for the estimation of discrete choice models. In *Swiss transport research conference.*

Bovy, P. H. L., Bekhor, S., & Prato, C. G. (2008). The factor of revisited path size: Alternative derivation. *Transportation Research Record*, *2076*(1), 132-140.

Dial, R. B. (1967). Transit pathfinder algorithm. *Highway Research Record*(205).

Elizalde-Ramírez, F., Nigenda, R. S., Martínez-Salazar, I. A., & Ríos-Solís, Y. (2019). Travel plans in public transit networks using artificial intelligence planning models. *Applied Artificial Intelligence*, *33*(5), 440-461.

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, *57*(3), 238–247.

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.

Nahmias-Biran, B., Han, Y., Bekhor, S., Zhao, F., Zegras, C., & Ben-Akiva, M. (2018). Enriching activity-based models using smartphone-based travel surveys. *Transportation Research Record*, *2672*(42), 280-291.

Prato, C. G. (2009). Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, *2*(1), 65-100.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276.

Tomhave, B. J., & Khani, A. (2022). Refined choice set generation and the investigation of multi-criteria transit route choice behavior. *Transportation Research Part A: Policy and Practice*, *155*, 484–500.

Yao, R., & Bekhor, S. (2020). Data-driven choice set generation and estimation of route choice models. *Transportation Research Part C: Emerging Technologies*, *121*, 102832.