Toward explainable clustering for distinguishing traveler profiles

Leonor Gomes*[1], Shadi Sharif Azadeh[1], Serge P. Hoogendoorn[1], and Gonçalo Homem de Almeida Correia[1]

[1]Transport and Planning Department, Delft University of Technology, Netherlands

Short summary

This study proposes a novel interpretable clustering framework to identify traveler profiles that can support the design of sustainable transport policies. The approach introduces an unsupervised tree-structured clustering method that explicitly incorporates contextual feature subsets into the clustering process. By organizing features into predefined subsets and assigning them to successive layers of a tree, the method enables domain knowledge to guide how traveler profiles are constructed. At each node, the next clusters are generated using agglomerative hierarchical clustering with distance measures suited to numerical, categorical, or mixed data. A genetic algorithm is then used to explore and optimize the space of possible tree structures by maximizing a weighted average silhouette score. Applied to data from the Netherlands Mobility Panel, the framework produces traveler profiles defined by travel behavior and sociodemographic characteristics while maintaining interpretability in the clustering process.
Keywords: Traveler profiling; Clustering; Multi-modal transport; Genetic algorithms; Interpretability

1 Introduction

Reducing car dependency is a key objective for many cities aiming to address congestion, emissions, and urban livability. Achieving this transition requires policies that encourage shifts toward more sustainable transport modes. Designing effective interventions, however, depends on understanding the heterogeneity of traveler behavior and identifying meaningful traveler profiles.

Clustering methods provide a data-driven way to group individuals with similar characteristics and identify traveler profiles. However, traditional clustering algorithms can struggle to produce clear and meaningful profiles when datasets contain both numerical and categorical variables. Additionally, the resulting clusters are often difficult to interpret, limiting their usefulness for policy design where transparent traveler profiles are valued.

Decision trees offer an interpretable modeling framework due to their hierarchical structure and transparent decision rules (Basak & Krishnapuram (2005)). Existing approaches typically determine node splits using greedy approaches (Hu et al. (2024)), while more recent work formulates clustering trees as optimization problems to improve node splitting (ICOT proposed by Bertsimas et al. (2021); XClusters proposed by Hwang & Whang (2023)).

This study proposes a clustering approach based on an unsupervised tree structure that separates feature subsets across different layers of the tree. At each layer, node splits are determined using only the features assigned to that layer, allowing contextual information and domain knowledge to guide the construction of traveler profiles. These node splits are generated using agglomerative clustering, which defines a discrete search space. By combining agglomerative clustering with the Gower distance, the method can handle mixed numerical and categorical data. Then a genetic algorithm is applied to optimize the tree structure, looking to maximize the silhouette score.

Unlike the ICOT framework, which determines node splits by directly optimizing single-feature decision rules through mixed-integer optimization, the proposed approach generates node partitions using agglomerative clustering. Similarly, the XClusters method derives clusters using k-medoids and subsequently trains a decision tree to explain them, whereas, in the proposed methodology in this paper, clustering is integrated into the tree construction, and the overall tree structure is optimized using a metaheuristic.

The contributions of this study are twofold: (1) a method for constructing interpretable traveler profiles that incorporates feature context and domain knowledge, and (2) a clustering framework

that employs agglomerative clustering for node splitting while optimizing the tree structure through a metaheuristic.

2 Methodology

This study proposes a tree-structured unsupervised clustering method. Starting from an initial set of features, feature subsets are defined using domain knowledge. For example, a subset may consist of features describing transport mode usage. Each feature subset defines a layer of the tree. In addition, each subset is assigned a data type, allowing the method to distinguish between numerical, categorical, and mixed data subsets.
The algorithm operates sequentially across the layers of the tree. At the first layer, the sample is clustered using only the feature subset associated with that layer. The clusters obtained are treated as nodes of the tree. In the second layer, each node is clustered independently using the second feature subset. This mechanism is repeated iteratively for each subsequent layer. The clusters obtained in the final layer form the final clustering output.
At each node, hierarchical agglomerative clustering can be applied. It is also possible that no clustering is done in a node. The number of clusters at each node is bounded by a predefined maximum parameter. Each node is associated with two decision variables: the number of clusters generated at that node and the linkage method used for the agglomerative clustering.
For categorical or mixed feature subsets, clustering uses the Gower distance metric (Gower (1971)) with single, average, or complete linkage. For numerical subsets, Euclidean distance is used and Ward linkage is additionally allowed. Figure 1 illustrates an example of the structure of a tree solution. For node i in layer l, there is an argument that will specify how many clusters the node will be separated into and which linkage method will be used. For the first case, "A-2-w" means the sample will be divided in two clusters with the agglomerative algorithm and ward linkage method.

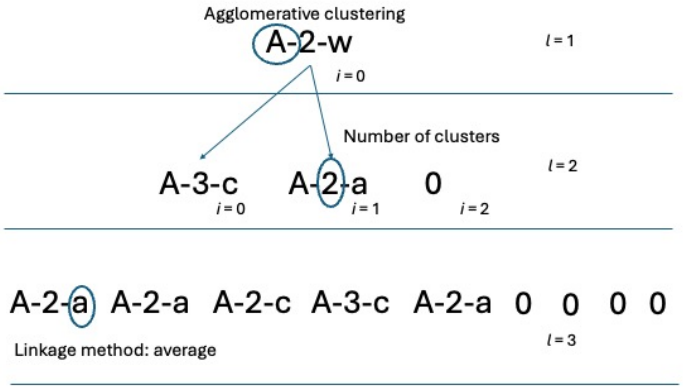


Figure 1: Tree Representation

A candidate tree solution must satisfy consistency constraints on the number of clusters and the linkage methods applied at each layer. If a layer contains categorical variables, Ward linkage cannot be used.
A tree-based clustering solution is evaluated by assessing the quality of the clusters generated at each node. To obtain an overall evaluation of the tree, the clustering metric is computed at every step and averaged into a single aggregate score. As no ground-truth labels are available for the profiles in this study, an internal validation metric is required. We use the silhouette score, a widely applied internal clustering metric that evaluates how well each observation fits within its assigned cluster compared to other clusters. It captures both intra-cluster cohesion and inter-cluster separation and is robust to clusters with non-spherical shapes and moderate outliers (Arbelaitz et al. (2013)).
The cluster-level silhouette scores are computed after each clustering step. Let $\mathbf{X}_{l,i}$ denote the partition of travelers obtained at node i in layer l. The silhouette score for this clustering is defined as

$$s_{l,i} = \text{Sil}(\mathbf{X}_{l,i}) \tag{1}$$

where $s_{l,i}$ represents the silhouette score associated with node i in layer l.
The objective function maximizes the weighted average silhouette score across all clustering nodes in the tree. Let $a_{l,i}$ be a binary variable indicating whether clustering is applied at node i in layer l ($a_{l,i} = 1$ if clustering is performed, and 0 otherwise). Let $n_{l,i}$ denote the number of travelers contained in that node. The objective function is defined as

$$\max \frac{\sum_{l,i} n_{l,i} s_{l,i} a_{l,i}}{\sum_{l,i} n_{l,i} a_{l,i}} \tag{2}$$

where it calculates the average of the clustering quality at each node, and not the global cluster results.
To find the optimal tree solution, a genetic algorithm (GA) is employed. A metaheuristic approach is preferred over an exact method due to the complexity of the problem. The GA evolves a population of candidate trees by applying crossover and mutation operators to variables such as the number of clusters and the linkage method.
The algorithm starts with an initial population of trees and runs for a fixed number of generations. In each generation, offspring equal in size to the population are produced. One parent is selected randomly, while the other is chosen through tournament selection, where the best solution from a random subset of the population is selected.
Crossover generates two offspring trees, which may undergo mutation with a given probability. A steady-state strategy is used to form the next generation by retaining the best half of the combined set of current solutions and offspring. To avoid local optima, if no improvement occurs for a predefined number of generations mutation is applied to all individuals to increase exploration of the solution space.

3 Results and discussion

The data used in this study comes from the Netherlands Mobility Panel 2022 (MPN) (Hoogendoorn-Lanser et al. (2015)), a household panel survey that captures travel behaviour in the Netherlands. The dataset includes a three-day travel diary with detailed trip information and questionnaires on participants' sociodemographic characteristics. After data cleaning, the sample size is 2902 travelers. From the sociodemographic data, two variables were selected: income and the presence of children. Income was encoded as an ordinal variable (0–2) reflecting increasing income levels. The number of children was converted into a binary indicator for having children under 12. Travel behaviour variables were also included. These consist of the total distance travelled over the three diary days, the share of trips made by mode category (active modes, public transport, and car), and the main commuting mode reported by each participant. The combination of sociodemographic characteristics and observed travel behavior allows profiling individuals not only by what they do, but also by the constraints and resources shaping their choices (Dieleman et al. (2002)).
The dataset includes sampling weights designed to correct for biases in sample composition and ensure representativeness with respect to key sociodemographic characteristics of the Dutch population. These weights are not used during the clustering phase, which aims to identify travel patterns, but are applied in the post-clustering analysis to obtain population-representative estimates.
Based on these features, the first subset used in the first layer consisted of the distance traveled. The second subset included all transport mode-related features, namely the car ratio, public transport ratio, active mode ratio, and the main transport mode used for commuting. The final subset contained the sociodemographic features, specifically income and whether travelers had children. The subsets were ordered to reflect a logical progression in travel behavior: total distance traveled first captures overall mobility intensity, transport mode features second reveal habitual travel patterns, and sociodemographic features last provide contextual constraints. However, alternative orders can be explored and finetuned depending on the application.
To obtain preliminary results for the traveler profiles, the algorithm was run for 200 generations with a maximum of three clusters per node, a mutation probability of 50% and a tournament size of 5. For the first 50 generations, the population contained 50 trees to encourage exploration, while the remaining 150 generations used 20 trees to focus on exploitation.

The fitness convergence of the genetic algorithm is shown in Figure 2. The best fitness value was obtained before the 25th generation, suggesting that the search space can be explored efficiently or that convergence occurs prematurely. Further experiments on the genetic algorithm parameters are required to distinguish between these two effects.

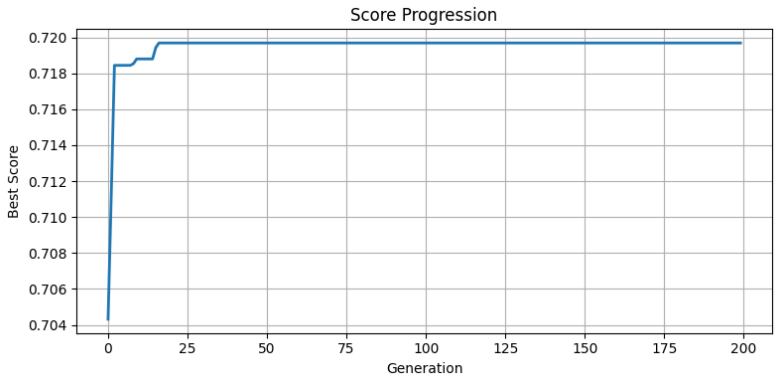


Figure 2: GA Convergence

The traveler profiles are summarized in Table 1. The tree structure in Figure 3 illustrates how the algorithm separated travelers at each node, with each node indicating the number of travelers it contains. In the first layer, the minimum and maximum travel distances are shown (e.g., d0.2–116 indicates a minimum of 0.2 kilometers in 3 days and a maximum of 116). The second layer displays the most common travel mode for each node, while the third layer divides travelers by sociodemographic characteristics.

Table 1: Results

	Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Indicators	Size	44	247	122	57	11	81	10	37	127	1156	62	14	31	820	83
Distance (kms / 3 days)	Mean	174	175.9	183.55	174.12	183.84	184.1	160.5	199.99	54.66	51.05	55.9	62.76	61.51	27.42	28.69
Income	Mean	1.34	0.85	2	1.25	1.27	0.74	1.4	2	1.22	0.99	0.71	1.21	2	0.98	1.06
Kids	Yes	44	0	0	0	11	0	10	0	127	0	0	14	0	0	83
	No	0	247	122	57	0	81	0	37	0	1156	62	0	31	820	0
Car Ratio	Mean	0.93	0.92	0.91	0.34	0.33	0.66	0.36	0.67	0.87	0.87	0.15	0.36	0.27	0.08	0.11
Active Ratio	Mean	0.04	0.05	0.06	0.06	0.05	0.14	0.43	0.12	0.12	0.11	0.10	0.07	0.16	0.88	0.88
PT Ratio	Mean	0.03	0.03	0.03	0.60	0.62	0.19	0.21	0.21	0.01	0.01	0.74	0.57	0.57	0.04	0.01
Main Transport Commute	Active	0	0	0	0	0	81	10	37	17	200	0	0	0	747	72
	Car	44	247	122	0	0	0	0	0	110	956	0	0	0	53	8
	PT	0	0	0	57	11	0	0	0	0	0	62	14	31	20	3

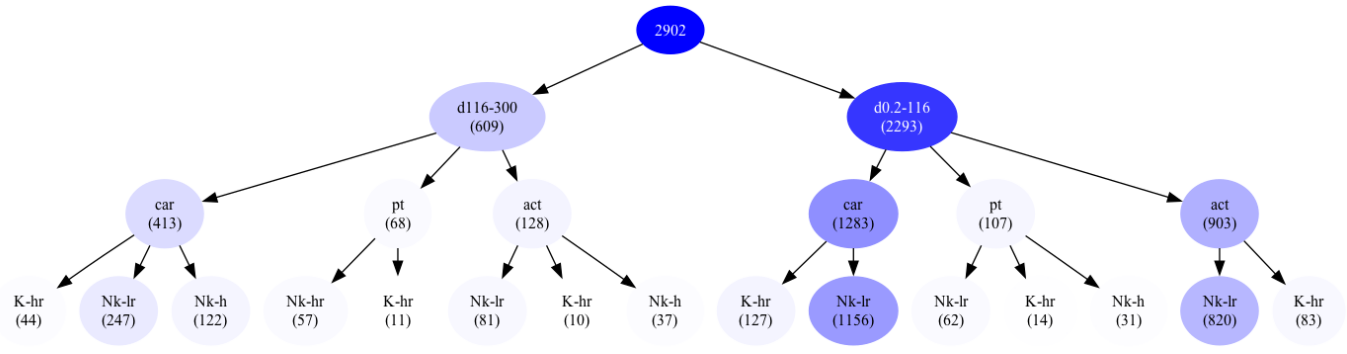


Figure 3: Clustering Tree

modes: pt - public transport; act - active modes; kids: K - kids; Nk - no kids; income: l - low; lr- lower; m - medium; hr - higher; h - high

The clusters identified can be found in table 2. The sampling weights have been applied to the clusters, and the corrected percentages are in the column "Dutch Pop". The percentages discussed below are sample-specific to demonstrate the separation made by the algorithm.
Clusters 10 (average distance travelled, car usage, no kids and below average income), 14 (average distance travelled, active mode usage, no kids and below average income), and 2 (long distance

Table 2: Overview of traveler profiles

Cluster (Size %)	**Distance**	**Main Mode**	**Kids**	**Income**	**Dutch Pop (%)**
1 (1.52)	Long	Car	Yes	Above average	2.89
2 (8.51)	Long	Car	No	Below average	7.26
3 (4.2)	Long	Car	No	High	3.61
4 (1.96)	Long	PT	No	Above average	1.85
5 (0.4)	Long	PT	Yes	Above average	0.57
6 (2.79)	Long	Active	No	Below average	2.25
7 (0.3)	Long	Active	Yes	Above average	0.61
8 (1.27)	Long	Active	No	High	1.16
9 (4.38)	Average	Car	Yes	Above average	7.25
10 (39.8)	Average	Car	No	Below average	34.99
11 (2.14)	Average	PT	No	Below average	2.53
12 (0.5)	Average	PT	Yes	Above average	0.75
13 (1.07)	Average	PT	No	High	0.76
14 (28.3)	Average	Active	No	Below average	28.21
15 (2.86)	Average	Active	Yes	Above average	5.32

travelled, car usage, no kids and below average income) contain the largest shares of travelers, representing 39.8%, 28.3%, and 8.5% of the sample, respectively. In contrast, clusters 7 (long distance travelled, active modes usage, kids and above average income), 5 (long distance travelled, public transport usage, kids and above average income), and 12 (average distance travelled, public transport usage, kids and above average income) contain the smallest shares, accounting for 0.3%, 0.4%, and 0.5% of travelers.

From Figure 3, the first layer separates travelers into two distance intervals: from 0.2 to 116 km (measured over three days) and from 116 to 330 km. All nodes originating from each branch therefore correspond to travelers whose total distance traveled falls within these respective intervals. In this sample, most travelers belong to the shorter distance interval, representing 79% of the population and indicating that the majority travel short to average distances. The second layer identifies the dominant transport mode within each distance interval. Based on the transport mode ratios and the reported main commuting mode, three main categories emerge: car, public transport, and active modes. In clusters 6 and 8, the algorithm prioritizes the reported main commuting mode. This situation can occur when travelers indicate an active mode as their main commuting mode but did not record any commuting trips during the three-day observation period. Furthermore, differences in modal distribution can be observed between the two distance intervals. For the shorter distance interval, the proportion of travelers primarily using active modes is higher (39.4%) than in the longer distance interval (21%). The opposite pattern is observed for public transport users: 4.7% in the shorter distance interval compared with 11.2% in the longer distance interval. Car use remains the most prevalent mode in both intervals, accounting for 55.9% of travelers in the shorter distance group and 67.8% in the longer distance group.

The final layer separates the nodes based on sociodemographic features. The algorithm primarily uses the presence of children as the main splitting criterion. However, given the previous separations by distance and transport mode, certain patterns related to income level can also be observed. In particular, clusters that include travelers with children tend to have higher average incomes. When comparing the distance intervals, travelers with children represent 9.8% of the shorter-distance group and 10.7% of the longer-distance group. For shorter distances, the proportion of travelers with children does not vary substantially across transport modes. In this interval, travelers with children represent 9.9% of car users, 13.1% of public transport users, and 9.2% of active mode users. For longer distances, the differences across transport modes are more pronounced. Among car users, travelers with children represent 10.7%, while the share increases to 16.2% among public transport users and 28.9% among active mode users. Regarding income, for shorter distances, the algorithm identifies clusters with lower average income across all transport modes. In contrast, for longer distances, lower-income clusters are only identified among car users and active mode users. This means that clusters corresponding to public transport users in the longer distance interval are associated with higher average income levels. Lower income groups who have to travel long distances are not taking public transport. These profiles highlight distinct behavioral groups that

may respond differently to policies encouraging modal shift.
The results presented are preliminary, and further analysis is needed to strengthen the evaluation of the proposed methodology. The next steps for the results will include sensitivity analysis of key parameters and comparison with traditional clustering approaches, such as hierarchical agglomerative clustering and k-means.

4 Conclusions

This study proposed a clustering framework based on an unsupervised tree structure, where agglomerative clustering determines node splits and a genetic algorithm optimizes the tree structure. The framework enables the incorporation of domain knowledge by separating features into contextual groups across different layers of the tree. In addition, the genetic algorithm efficiently explores the space of possible tree structures and identifies high-quality solutions early in the optimization process.
The method identifies 15 traveler profiles characterized by distance range, dominant transport mode, presence of children, and income level. The results show that transport mode usage varies across distance ranges: car use dominates both shorter and longer distances, active modes are more prevalent for shorter trips, and public transport users are more common in the longer-distance range. In the sociodemographic layer, the presence of children emerges as the primary splitting variable, with clusters containing travelers with children generally associated with higher income levels. These interpretable traveler profiles can support the design of targeted mobility policies aimed at encouraging shifts toward more sustainable transport modes.

Acknowledgements

This work has been funded by the European Union under the Horizon Europe grant 101139678 (metaCCAZE). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

References

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013, January). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243–256.

Basak, J., & Krishnapuram, R. (2005, January). Interpretable hierarchical clustering by constructing an unsupervised decision tree. *IEEE Transactions on Knowledge and Data Engineering*, *17*, 121–132.

Bertsimas, D., Orfanoudaki, A., & Wiberg, H. (2021, January). Interpretable clustering: an optimization approach. *Machine Learning*, *110*(1), 89–138.

Dieleman, F. M., Dijst, M., & Burghouwt, G. (2002, March). Urban Form and Travel Behaviour: Micro-level Household Attributes and Residential Context. *Transportation Research Procedia*, *39*, 507–527.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857–871.

Hoogendoorn-Lanser, S., Schaap, N. T. W., & OldeKalter, M.-J. (2015, January). The Netherlands Mobility Panel: An Innovative Design Approach for Web-based Longitudinal Travel Data Collection. *Transportation Research Procedia*, *11*, 311–329.

Hu, L., Jiang, M., Dong, J., Liu, X., & He, Z. (2024, June). Interpretable Clustering: A Survey. *ACM Computing Surveys*, *58*(8), 1–21.

Hwang, H., & Whang, S. E. (2023, June). XClusters: Explainability-First Clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(7), 7962–7970.