How Rational are Travellers in Zurich? Exploring Public Transport Trips from a Shortest Path Perspective

Mariana A. Costa

Francesco Corman

STRC conference paper 2024 April 25, 2024

STRC | 24th Swiss Transport Research Conference
Monte Verità / Ascona, May 15-17, 2024

How Rational are Travellers in Zurich? Exploring Public Transport Trips from a Shortest Path Perspective

Mariana A. Costa
IVT
ETH Zürich
CH-8093 Zurich
mariana.de-almeida-costa@ivt.baug.ethz.ch

Francesco Corman
IVT
ETH Zürich

April 25, 2024

Abstract

This paper uses unlabelled GPS tracking data, enriched by fusion with Automatic Vehicle Location (AVL) data, to assess the rationality of individuals' travel choices in the city of Zurich, Switzerland. By examining thousands of public transport journeys, we introduce trip metrics normalized by the respective shortest path, providing an intuitive method to compare trips across diverse users and attributes. According to the patterns encountered, we further classify trips into five distinct groups: i) the 'as fast as possible'; ii) too early or too late for connections; iii) walking-prone; iv) sitting-prone; v) do not mind a bit of walking. By employing a tree-boosting algorithm, we demonstrate that a statistical model can learn to distinguish well between the five groups of trips. The findings suggest that our approach is promising for aiding planners in optimizing routes and schedules to enhance efficiency and meet passenger demand. This research also provides valuable insights into user behaviour by showing that individuals exhibit a mix of different trip patterns, rather than a single predominant behaviour, when choosing their trips. These insights on individual behaviour heterogeneity can contribute to data-driven improvements in public transportation systems.

Keywords

GPS Tracking data; Public Transport; Route Choice; Shortest Path; Rationality of Travellers; Trip Clustering; Behavioural Patterns

1 Introduction

In recent years, research on travel behaviour and mobility patterns has greatly benefited from massive sources of data, such as GPS tracking and smartcard data. The challenge remains on how to use and analyse these data to better understand how socio-demographics, trip attributes, network infrastructure, and personal preferences interact in the complex travellers' choice process. Modelling and extracting features that have a significant influence on this complex travel choice process is pivotal for travel demand prediction and improved travellers' satisfaction. Previous research has shown that human trajectories have a high degree of temporal and spatial regularity, with a high probability of returning to a few highly frequented locations (Gonzalez *et al.*, 2008) and that most individuals travel short distances in a well-localized and finite neighborhood, exhibiting a high potential of predictability (Song *et al.*, 2010).

This repetitive nature of human mobility patterns has motivated many researchers to depict travellers' behaviour using clusters of individuals with similar characteristics towards their travel decision-making process and/or activity patterns. While many works focus on human or activity-based clustering, emphasizing groups of individuals, trajectory-based clustering primarily addresses route choice and travel decisions, capturing movement patterns and paths. In both cases, however, measuring multidimensional similarity is a critical problem in applying clustering techniques (Zhai *et al.*, 2019). Ultimately, one seeks to quantify the interplay between the regular and thus predictable and the random and thus unforeseeable (Song *et al.*, 2010). Although human mobility patterns are known to be repetitive, they are also highly dependent on temporal variables (e.g., the day of the week (Thuillier *et al.*, 2017)), as well as the activity-related variables (e.g., time taken for different activities, characteristics of the activities, purpose, or sequencing of the event (Krause and Zhang, 2019)).

Clustering groups of individuals or activities may not only require multiple travel features, e.g., the sequence of location and time, duration, trip purpose, trip mode, accompanying persons, etc., but also related socio-demographic characteristics, making it challenging from the perspective of data acquisition (Zhai *et al.*, 2019). In contrast, trajectory-clustering usually relies on rich trajectory datasets, such as those acquired from passive GPS tracking, mainly focusing on the spatiotemporal dimensions. From a route choice perspective, research has shown that the underlying user behaviour varies by trip purpose, and individual characteristics influence route choices even for the same trip purposes (Dalumpines and Scott, 2017). Nevertheless, to describe route choice behavior, identify potential navigation problems, design more readable cities, and provide understandable travel information, it is important to understand how variations in urban wayfinding behavior relate to everyday travel patterns Sivalingam *et al.* (2024). Hence, grouping trips instead of individuals is better aligned with the goal of understanding the overlap between the actual and the "optimal" (from a service provider point of view) journeys. This work delves into the idea of trajectory-clustering. Still, instead of the trajectories themselves, the idea is to group trips together based on route-specific attributes normalized by their shortest path counterpart. We further limit the analysis to public transport (PT) trips, with the objective of measuring PT route choice efficiency based on simple, readily available metrics in the context of a network well-known for offering reliable, high-frequency connections. Specifically, this paper adds to the state of the art:

i. it proposes an approach relying only on GPS tracking and AVL data to characterize and distinguish different groups of trips with respect to their performance compared to shortest path metrics. By choosing this research path, we can leverage passive GPS tracking data and offer behavioural insights even when complementary data, e.g., socio-demographics or trip purpose, are not available.

ii. it defines and summarizes trip behaviour efficiency/rationality in terms of five distinct groups in the context of spatiotemporal patterns. By defining the trip choice process in terms of proximity to the shortest path trip, we can make behavioural inferences on the rationality of users using PT.

This paper uses travel diaries collected by a smartphone application called *ETH-IVT Travel Diary* consisting of 2909 trips of 172 users in the city of Zürich (Switzerland). The application allowed (continuous) passive tracking, and activities, trips, and modes were identified through a mode detection algorithm, as described in Marra *et al.* (2019). Further information and some descriptive statistics on the survey questionnaire can be found in the Appendix. The paper continues with a literature review. Then, Section 3 reports the methodology. Section 4 presents the clustering of trips and analyses the results. Section 5 concludes the paper.

2 State of the art

A recent literature review indicated that travel behavioural research using GPS data has been broadly investigated in recent years (Sivalingam *et al.*, 2024). The results demonstrated that GPS, which offers precise, time-stamped location data, is among the most important technological advancements to address the shortcomings of conventional travel surveys since the late 1990s and early 2000s. Compared with sensing data, survey data is disadvantaged by high cost, low frequency, and small sample size. On the other hand, because it often comes with socioeconomic and demographic information, survey data provides rich information for exploring differences underlying human activity dynamics (Jiang *et al.*, 2012). Hence, many works combine some sort of passive tracking (GPS, smartcard, mobile phone data, etc.) with socio-demographics (e.g., questionnaires or official surveys) for better understanding travel behaviour.

Clustering methods are generally used to exploit human mobility patterns ranging from spatiotemporal trajectories to trip characteristics (Zhou *et al.*, 2021), and more recently, there has been a lot of works centred on activity patterns, see table 1 for some an overview of the literature on clustering techniques employed for extracting trajectory, activity or user patterns.

Some other studies reinforce the importance of the spatial/temporal features chosen in terms of passenger behaviour by using other methods. Song *et al.* (2010) measure the entropy of individuals' trajectory using mobile phone data, and find high predictability and regularity of users' daily mobility. Ortega-Tong (2013) also explores the similarity in travel patterns from riders with smart cards combined with socio-demographic characteristics to identify clusters with similar structures. Carrel *et al.* (2013) conclude that departure regularity is the most important path-specific feature, and they show that PT passengers gradually adapt upon changes to departure reliability and headway lengths, although for short headways this adaptation consists mostly of a stochastic behaviour. Kusakabe and Asakura (2014) use smart card data to analyse behavioural features to classify trip purpose by utilising a naive Bayes classifier. For bus ridership, Kim *et al.* (2017) introduce a metric called "stickiness index" to classify users according to the regularity in which they choose their routes, which relates to the frequency of similar routes. Goulet-Langlois *et al.* (2017) hypothesize that the order in which an individual engages in trips and activities is an important characteristic of travel behaviour, so they propose an approach to measure the regularity of travel behaviour based on the order in which travel events are organised over time in travel sequences. They conclude that travel regularity may follow atypical patterns which are not captured by either periodicity-based methods or activity-based models.

All these studies take advantage of data available and apply methods that enable inference about travellers' behavioural characteristics. The need to complement passive data with

Table 1: Overview of literature on clustering techniques employed for extracting trajectory, activity or user patterns

Authors (Year)	(Main) Dataset	Location	Clustering Method	Similarity Measure	Objective
Joh *et al.* (2001)	2-day activity diaries	Hendrik-Ido-Ambacht and Zwijndrecht, Netherlands	Hierarchical Clustering	Trajectory-based multi-dimensional sequence alignment	Compare performance of the proposed multi-dimensional sequence alignment methods for classifying travel patterns with traditional distance-based and signal-processing approaches
Jiang *et al.* (2012)	Activity-based travel survey	Chicago, USA	K-means	Euclidean distance	Explore the daily activity structure and its variation, and cluster individual behaviour with further comparison with socio-demographics
Zheng *et al.* (2012)	Geotagged photos	Paris, London, San Francisco and New York	Hierarchical Clustering	Similarity of tourist travel routes based on longest common subsequence of visited locations	Analysis of tourist movement trajectories based on a Markov chain framework
Ma *et al.* (2013)	Smart card data	Beijing, China	DBSCAN, K-means++	Euclidean distance	Fast data-mining procedure that models the regular travel patterns of transit riders
Goulet-Langlois *et al.* (2016)	Smart card data	London, UK	Hierarchical Clustering	Euclidean distance	Identify clusters of users with similar activity sequence structures
Ma *et al.* (2017)	Smart card data	Beijing, China	Technique for order preference by similarity to an ideal solution (TOPSIS)	Euclidean distance	Investigate spatial and temporal travel patterns in Beijing
Zhang *et al.* (2017)	Geotagged tweets	Northern Virginia, USA	Sequential model-based clustering	Gaussian distribution	Extraction of travel behaviour using social media location data and comparison to household survey data
Wang *et al.* (2018)	GPS taxi trajectory data	Wuhan, China	Hierarchical Clustering	Edit Distance	Detecting anomalous taxi trajectories
Thuillier *et al.* (2017)	Call detail records (CDR) mobile phone data	Paris suburban area, France	K-means	Hamming distance	Characterize human mobility patterns based on cellular data with further validation with a National Census
Zhai *et al.* (2019)	Household trip survey data	Puget Sound Region, USA	Affinity propagation (AP) clustering	Augmented space-time-weighted edit distance	Measuring similarities between human activities patterns
Zhou *et al.* (2021)	Household trip survey data	Nanjing, China	Markov-chain based mixture model clustering		Clustering of travellers based on activity patterns
Costa *et al.* (2023)	GPS Tracking data	Zurich, Switzerland	DBSCAN	Euclidean Distance	Clustering of travellers (commuters) based on trip patterns

socio-demographics relates to the inherently dynamic nature of human behaviour, which from a trip planning perspective may involve many features not readily available with passive tracking, such as trip purpose and characteristics of activities (daily routine arrangement, activity schedule, relevance, etc.). Nevertheless, raw trajectory data such as GPS tracking data (which provides very low-level information, comes with relatively low quality and a non-systematic sampling rate), fused with AVL (Automatic Vehicle Location) data (which describes the actual PT supply) is a promising solution to expand and develop the investigation of spatiotemporal patterns pertaining urban mobility, especially those directly related to route choice efficiency, which are further investigated in this paper. This work differs from the others by presenting a methodology using complete trip itineraries obtained from GPS tracking data to distinguish different groups of trips based on how they deviate from the shortest path trip. Based on readily available spatiotemporal trip metrics and further normalization of those metrics by the shortest path attribute, we aim to classify different groups of trips and behaviours towards trips that measure how “logically" users make decisions.

The next sections utilise the trips identified by the mode detection algorithm in the *ETH-IVT Travel Diary* survey. The mode detection algorithm described in Marra *et al.* (2019) derives travel diaries from GPS data. The algorithm identifies, for each user, activities (done in a single location) and trips (movements between activities). Afterward, each trip is divided into stages, and the mode for each stage is identified (walk, private or public transport). For each public transport stage, the algorithm also identifies the vehicle, line, departure and arrival stops and times. Without going into the details, the main criteria to detect the mode are a low speed to identify walks, and a comparison with AVL data to identify public transport stages. By comparing the path of a public transport vehicle (from the AVL data) with the path of a user (from GPS data), it is possible to detect the vehicle used. Finally, a private stage is identified by exclusion and can be further distinguished into car or bike, if needed. Overall, the mode detection algorithm has an average accuracy of 86.14% and has been validated in previous works on the same dataset (Marra and Corman (2020); Marra *et al.* (2022)), identifying realistic mode share and estimation of route choice models.

3 Methodology

The proposed methodology is a three-step procedure which incorporates machine learning methods and statistical analysis of the results. The first step is the extraction of route specific attributes, which are normalized to account for observations (trips) with different lengths, times and transfer patterns. The second step is the clustering of trips based on these route specific attributes and considering a Hierarchical Clustering approach. To check the quality of the clusters, we employ some known methods, such as the Silhouete score, elbow score and PCA visualization. The third and final step fits a tree-based model and performs statistical analysis on the results, including measuring the impact of variables through partial dependence plots (PDPs). The next subsections present a detailed description of each step.

3.1 Route Specific Attributes

The normalized route specific attributes are expected to capture the overall characteristics of the routes in a way that they can be compared to each other, not only in terms of the scale, but also from the standpoint of a standard and logical perspective: how much they deviate from their corresponding shortest path (i.e. the fastest realized trip). In order to accomplish that, two shortest paths can be considered: the timetable shortest path and the actual (realized) times shortest path. These shortest paths were derived in the work of Marra and Corman (2020), and are based on fusion of GPS tracking data (the trip information) with AVL information of PT. As explained in section 2, the algorithm had very high accuracy and was validated in previous works. Hence, for each trip, information on the actual path (route) of the user, could be contrasted with the timetable and actual shortest paths. For the remainder of this study, we move forward only with the actual shortest path from the PT provider. By inspection, it was found that many timetable trips suffered changes causing the comparison with the user path to generate unrealistic ratios. The actual shortest path could, therefore, more realistically capture the user behaviour without extreme outliers, and is used herein.

Each one of the six metrics presented is then normalized by the respective actual shortest path metric, so that each final metric is a ratio (proportion) of the shortest path counterpart, thus allowing for all trips to be compared to each other in terms of how much they deviate from the shortest path. These metrics are presented below.

3.1.1 Route Length Detour

The route length detour calculates, for each trip i, the percentage detour of the trip length in kilometres, D_i, to the corresponding shortest path length for that same trip, $D_{i,SP}$.

$$RouteLengthDetour_i = \frac{D_i - D_{i,SP}}{D_{i,SP}}$$

3.1.2 Route Directness Detour

The route directness detour calculates, for each trip i, the percentage detour of the number of transfers in the trip, NT_i, to the corresponding number of transfers in the shortest path for that same trip, $NT_{i,SP}$. If both the shortest path or the trip have no transfers a value of 0 is attributed to this indicator, otherwise, if any of the trips does not have a transfer, this indicator gets a value of "NaN" and is not further used.

$$RouteDirectnessDetour_i = \frac{NT_i - NT_{i,SP}}{NT_{i,SP}}$$

3.1.3 Route Transfer Detour

The route transfer detour calculates, for each trip i, the percentage detour of the transfer time (in minutes) of the trip, TR_i, to the corresponding transfer time in the shortest path for that same trip, $TR_{i,SP}$. If both the shortest path or the trip have no transfers a value of 0 is attributed to this indicator, otherwise, if any of the trips does not have a transfer, this indicator gets a value of "NaN" and is not further used.

$$RouteTransferDetour_i = \frac{TR_i - TR_{i,SP}}{TR_{i,SP}}$$

3.1.4 Route Time Detour

The route time detour calculates, for each trip i, the percentage detour of the trip total time in minutes, T_i, to the corresponding shortest path total time for that same trip, $T_{i,SP}$.

$$RouteTimeDetour_i = \frac{T_i - T_{i,SP}}{T_{i,SP}}$$

3.1.5 Route Walking Detour

The route time detour calculates, for each trip i, the percentage detour of the trip total wwalking time in minutes, W_i, to the corresponding shortest path total walking time for that same trip, $W_{i,SP}$.

$$RouteWalkingDetour_i = \frac{W_i - W_{i,SP}}{W_{i,SP}}$$

3.1.6 Route In Vehicle Detour

The route time detour calculates, for each trip i, the percentage detour of the trip total in vehicle (i.e. bus, train, tram) time in minutes, IV_i, to the corresponding shortest path total in vehicle time for that same trip, $IV_{i,SP}$.

$$RouteInVehicleDetour_i = \frac{IV_i - IV_{SP}}{IV_{i,SP}}$$

3.2 Hierarchical Clustering

The second step identifies routes with similar patterns in terms of their normalized route characteristic attributes by means of clustering. We run an agglomerative hierarchical clustering algorithm (Nielsen and Nielsen, 2016) using Ward's distance (see equation 1). In this clustering approach, each trip is defined as a vector of its normalized features and, at each step, two clusters merge given they provide the smallest increase in the combined error sum of squares (Ward's distance). This type of clustering starts from a single cluster (single trip or data point) and merges until all trips are combined into one or k clusters. Hence, it is necessary to choose a stop criterion (number of clusters, k).

$$d(u,v) = \sqrt{\frac{|v|+|s|}{|T|} \cdot d(v,s)^2 + \frac{|v|+|t|}{|T|} \cdot d(v,t)^2 + \frac{|v|}{|T|} \cdot d(s,t)^2} \tag{1}$$

where:

- $d(u,v)$ is the distance between clusters u and v,
- u is the newly joined cluster consisting of clusters s and t,
- v is an unused cluster in the forest,
- $T = |v| + |s| + |t|$, and $| * |$ is the cardinality of its argument.

The Hierarchical clustering technique can be visualized using a dendrogram, which is a tree-like diagram that records the sequences of merges or splits. Inspection of the dendrogram is helpful for determining a good cut-point for k, but other methods should also be considered, such as the silhouette score (see equation 2) and the elbow method (or k-elbow score). The silhouette score measures how close the samples are to their cluster centroids, and how far away the samples are to their neighboring clusters (Yao and Bekhor, 2020) . The score falls within the range of $[-1, 1]$, where 1 indicates that the data point is far away from its neighborhood cluster (a desirable characteristic), and -1 indicates that the data point has been assign to the "wrong" cluster. The elbow method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters; the "elbow point" corresponds to the number of clusters for which the rate of decrease in the metric sharply decreases, indicating that adding more clusters does not significantly decrease the WCSS (i.e. an "optimal" number of clusters is reached).

$$SilhouetteScore = \frac{1}{n}\sum_{j=1}^{n}\frac{b(x_j) - a(x_j)}{\max\{a(x_j), b(x_j)\}} \tag{2}$$

where:

- $a(x_j)$ is the average distance from data point x_j to all other data points in the same cluster,
- $b(x_j)$ is the minimum distance from data point x_j to all other data points in a different cluster.

Once clusters have been identified, statistical analysis on the classified patterns is performed to assess the different groups within the context of route choice efficiency/rationality.

3.3 Tree-based model

Following the clustering algorithm, a tree-based learning algorithm (LightGBM) with negative log-likelihood loss (equation 3) for multiclass classification is trained. This machine learning method minimizes the loss function by adjusting the parameters of the decision trees. This method has an advantage over traditional classification methods in modeling co-linear and interacting features, besides high predictive accuracy, since the principle of gradient boosting is to train a series of weak models and to combine their predictions to create a stronger prediction.

$$\text{Negative Log-Likelihood Loss} = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log(p_{ik}) \tag{3}$$

Where:

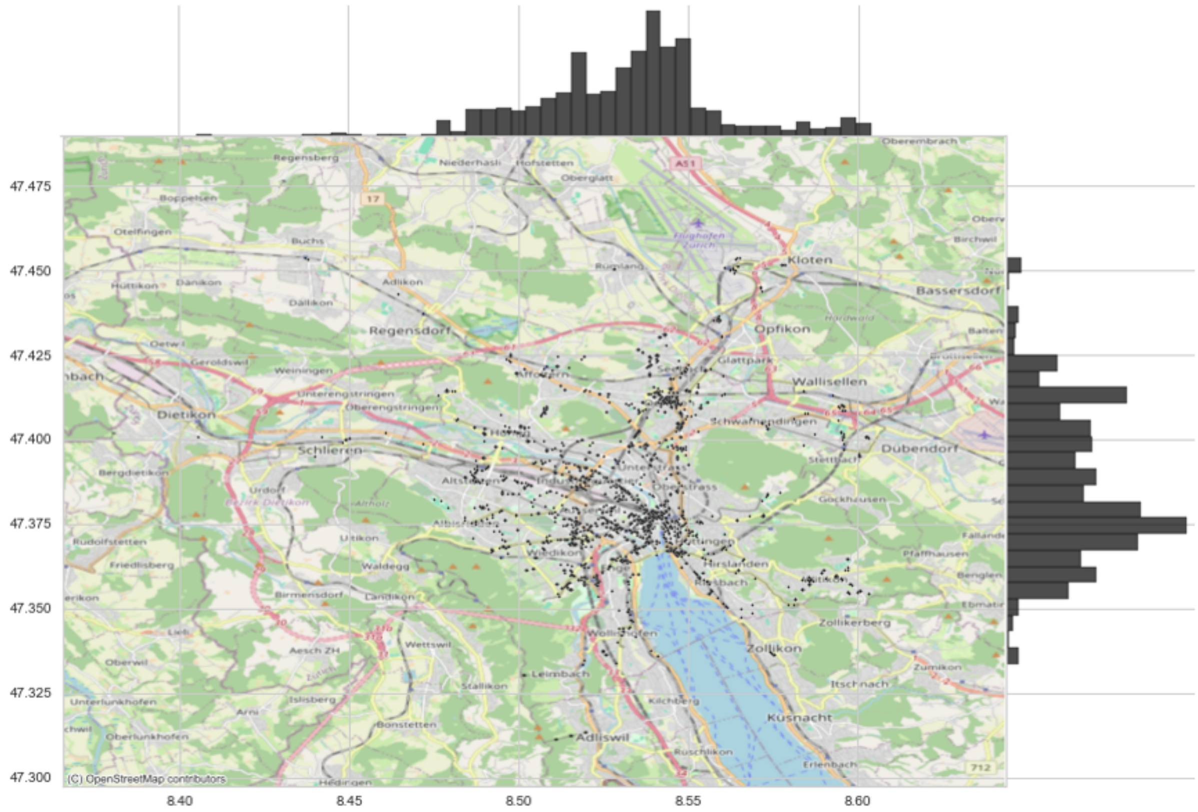
- N is the total number of samples,
- K is the total number of classes,
- y_{ik} is a binary indicator (0 or 1) of whether sample i belongs to class k,
- p_{ik} is the predicted probability that sample i belongs to class k.

While clustering helps identifying patterns and, in general, any structure within the data, by fitting a tree-based model on the clustered data we can further explore and understand the relationships between variables and how they contribute to the identified clusters. For instance, we can know how much each variable (e.g. walking route detour) impacts a specific cluster of trips. Hence, having a model is key for statistical analysis, predictions and further visualization, for example, in the form of partial dependence plots (PDPs), which help understanding how individual features influence predictions and provide insights into the relationships between features and the target variable within each cluster.

4 Results and Discussion

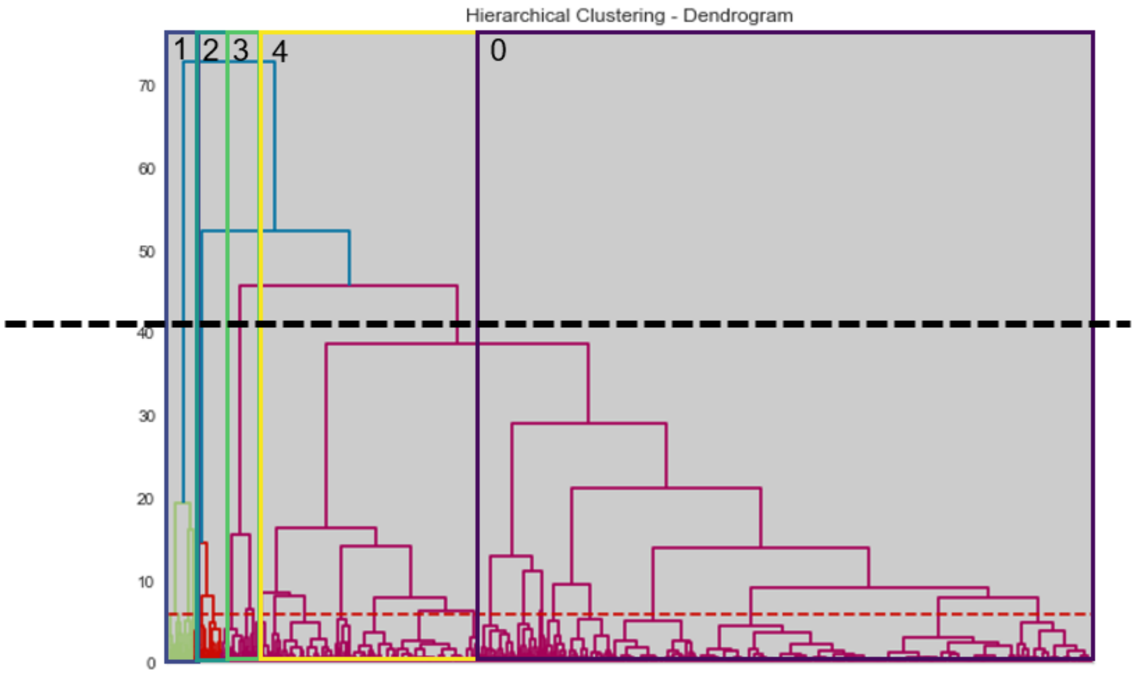
Figure 1 shows all origin and destinations pairs of the 2909 PT trips from all 172 users in the study. The histograms in the x- and y-axis indicate the geographical distribution of points, showing that trips with either origin or destination around the city centre are predominant.

Figure 1: Distribution of origin and destination points of trips.



Following the methodology, each trip is first represented by a set of route specific attributes normalized by the actual shortest path. The final feature set was selected based on the combined clustering approach and resulting accuracy of the tree-based model, after enumeration of the possible combinations of features. The four features selected were: *RouteTimeDetour, RouteTransferDetour, RouteWalkingDetour, RouteInVehicleDetour*. For the number of clusters, the methods highlighted in section 3 were used, with $k = 5$ clusters being the final choice. Figure 2 shows the dendrogram with the separation given the five clusters selected. The dashed black line shows the cut off selected, which resulted in the 5-cluster division. A trade-off between granularity and interpretability has to be taken into consideration: overfitting (increasing the number of clusters) can lead to more details being captured, but results may be hard to interpret and generalize. It is possible to check that cluster 0 (purple square) is the largest one, with also the highest variability. The other clusters are smaller and capture specific patterns of interest as we show in the upcoming analyses.

Figure 2: Dendrogram.



The Silhouette score (figure 3) and K-elbow score (figure 4) also indicated that setting $k = 5$ was a reasonable choice, given the interpretability vs. granularity trade-off and, in particular, the choice was suitable for the extraction and analysis of features of interest.

Figure 3: Silhouette score for different values of k.

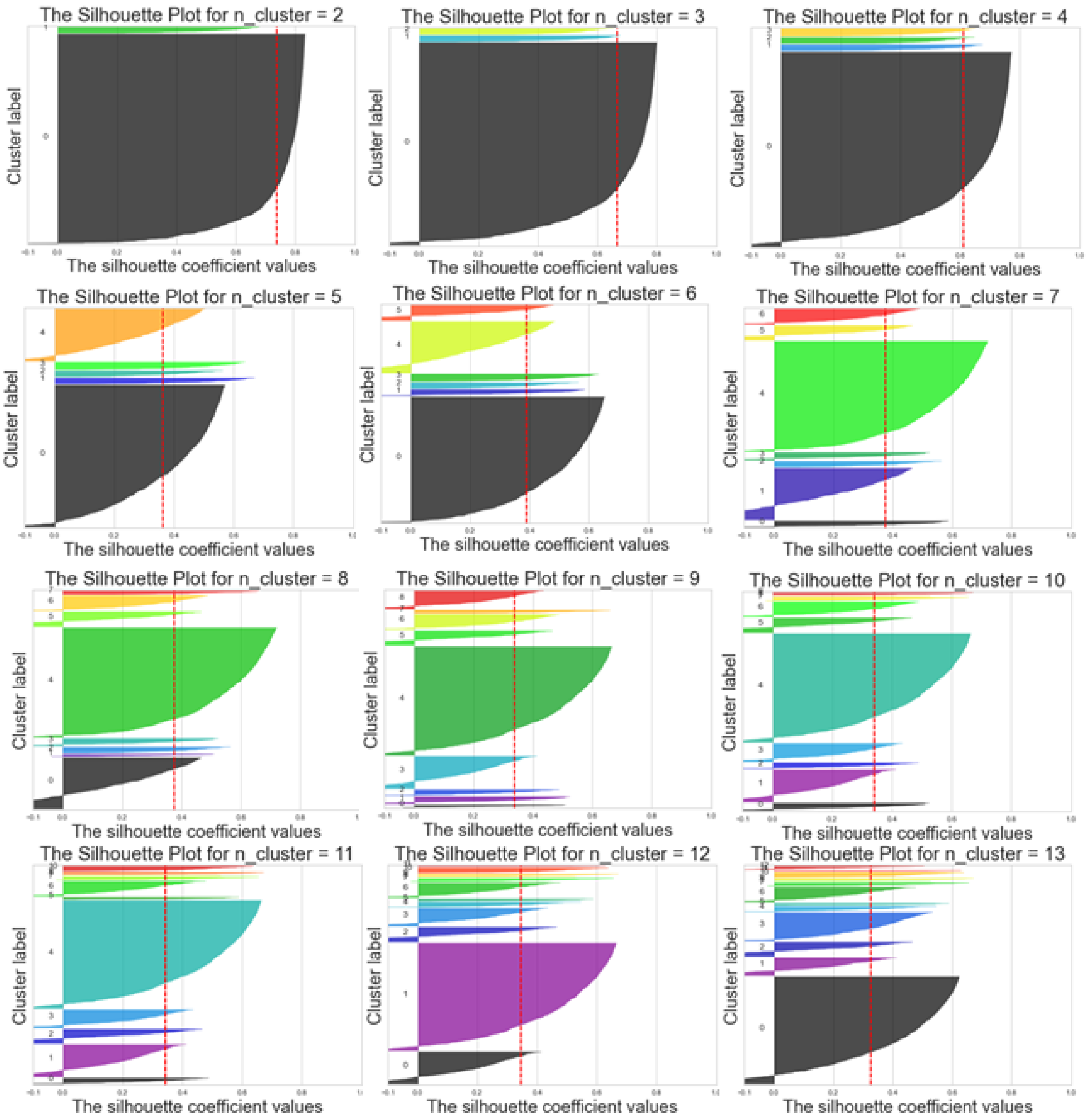
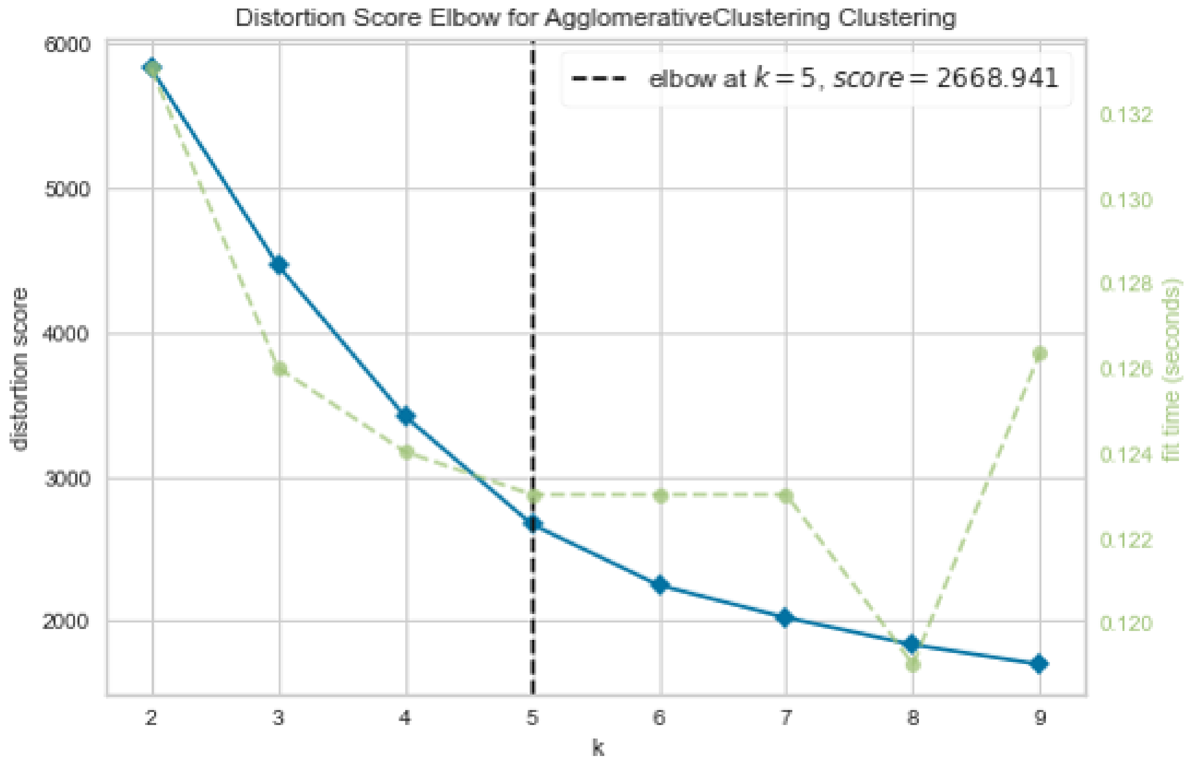
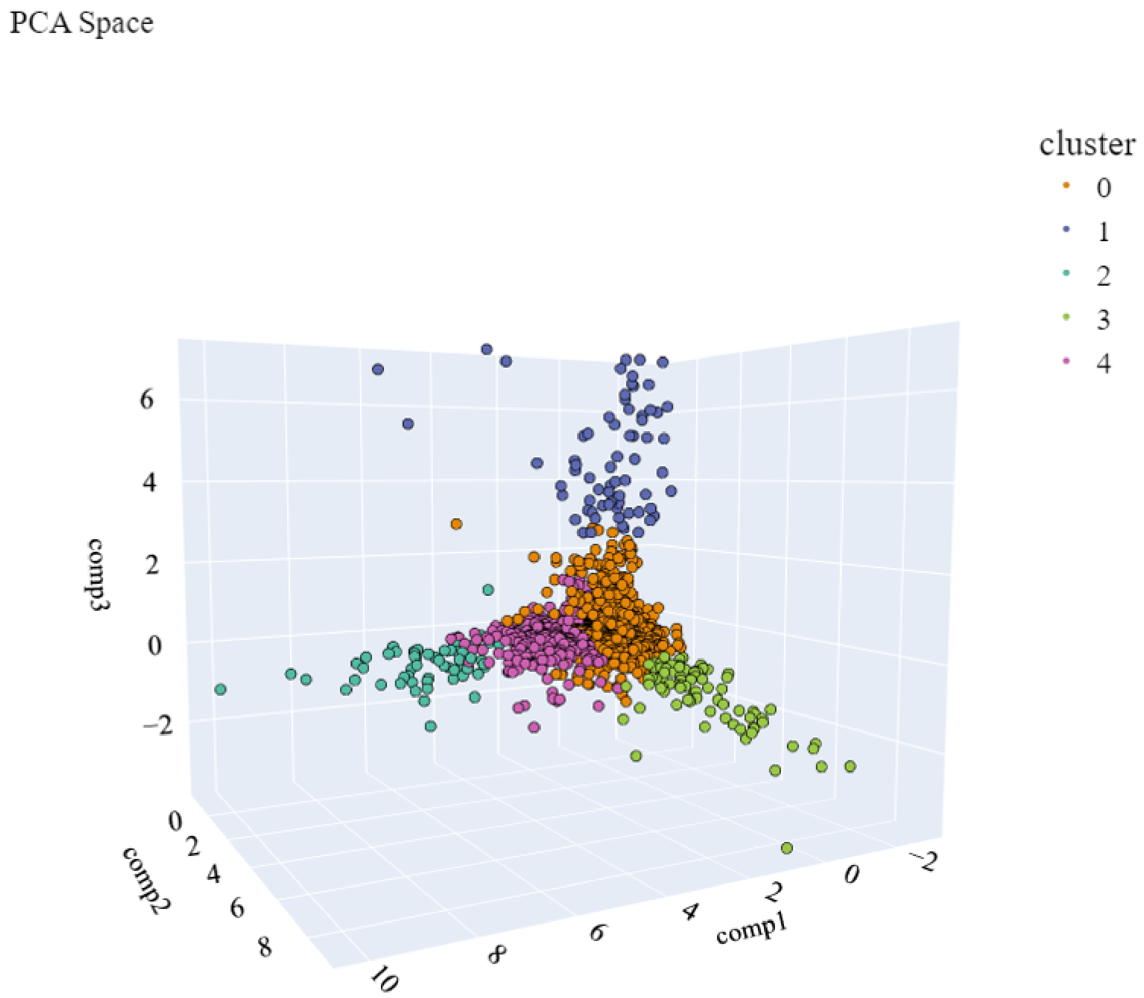


Figure 4: K-elbow score for different values of k.



The final clusters can be visualized in the PCA (Principal Components Analysis) space considering the first three principal components (see figure 5). This visualization is interesting as PCA transforms the original features into a set of uncorrelated variables (or the principal components). Hence, it helps in reducing the dimensionality of the data while retaining most of the variance, making it easier to visualize and interpret the results. The three first principal components shown in figure 5 account for approximately 95% of the accumulated variability in the data, revealing a good separation between clusters, which is desired for the following analyses.

Figure 5: PCA Space Visualization of Clusters.



From the five clusters obtained, some initial interpretation can be achieved with statistical visualization tools. Figure 6 shows scatter plots for the four features (*RouteTimeDetour, RouteTransferDetour, RouteWalkingDetour, RouteInVehicleDetour*) and the different data points coloured by the cluster number. From this figure, it is possible to identify patterns linked to each cluster depending on the feature. In particular, clusters 2 and 4 seem to be associated with higher walking times detours (at least two times over the standard shortest path), cluster 3 seem to be associated with higher in-vehicle detour times (also at least 2 times over the standard shortest path), and cluster 1 seems to be associated with higher transfer times (ranging from 4 to 10 times higher than the shortest path). Cluster

0, our default cluster, seems to be linked with values very close or even lower than the standard shortest path. To get a better sense of these numbers, figure 7 complements figure 6 by showing, for each feature (row), the five corresponding boxplots of the different clusters statistics. The colour scheme for both pictures (and throughout this paper) is kept the same for each cluster, for better visualization.

Figure 6: Scatter plots: clusters vs. features.

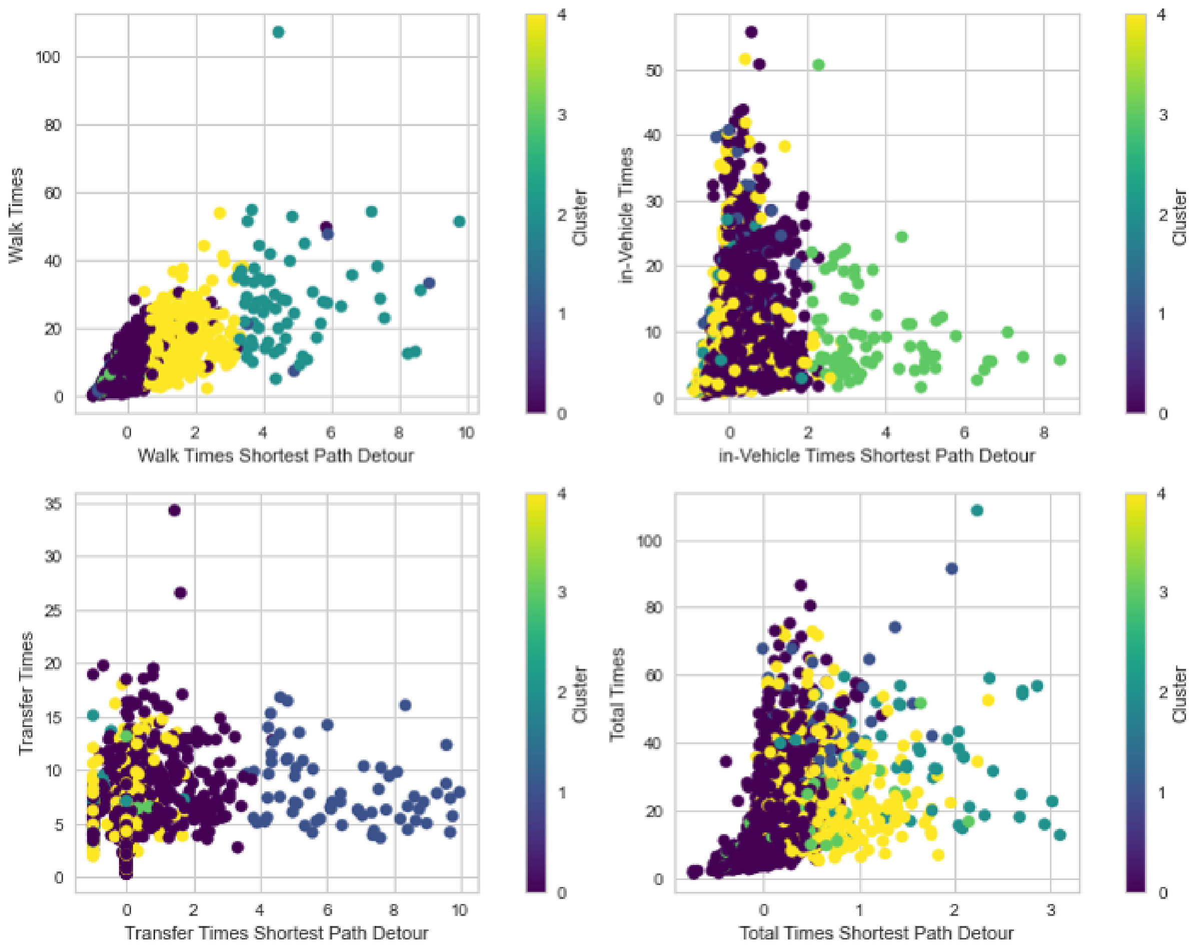
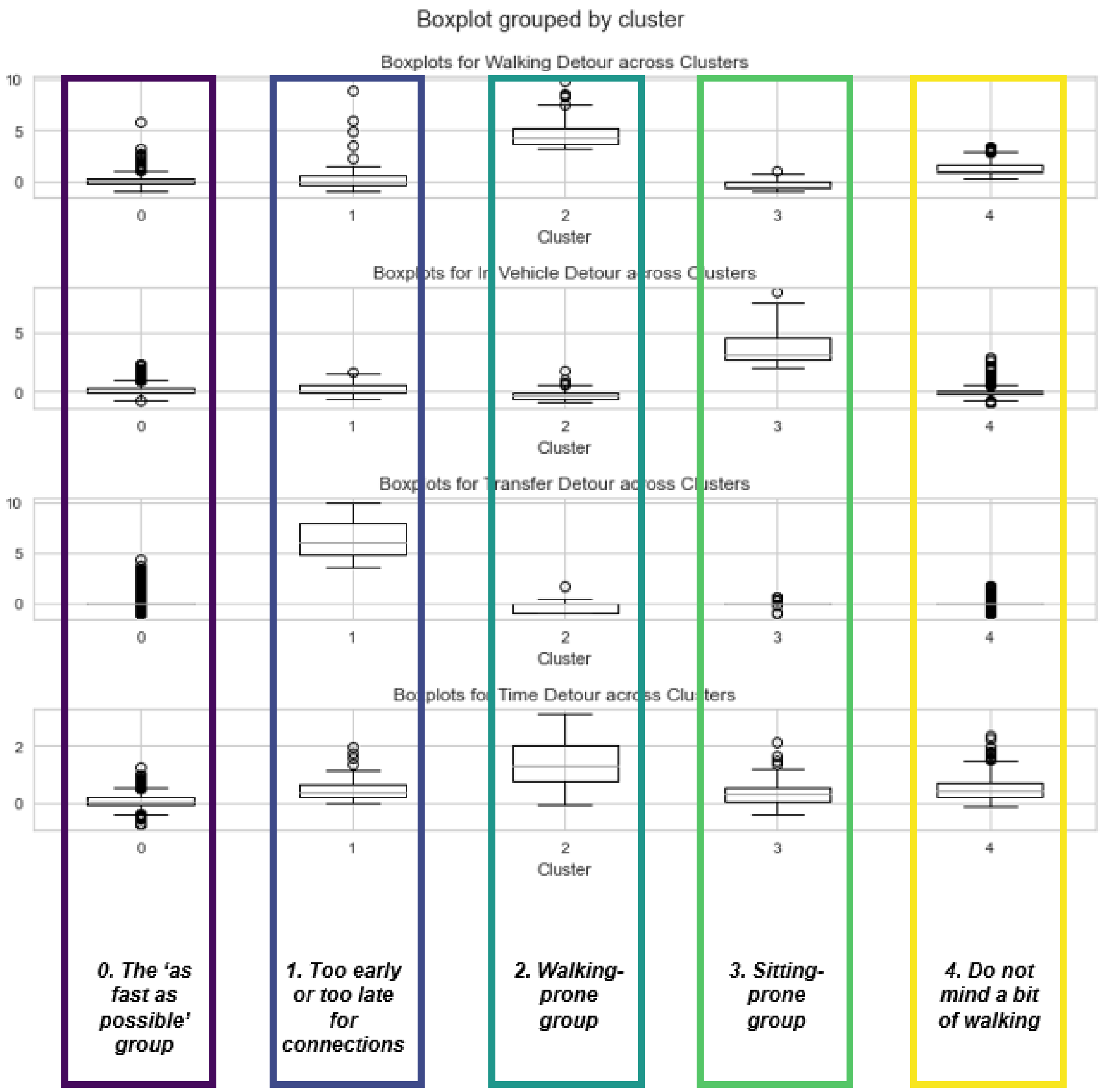


Figure 7: Histograms: clusters vs. features.



Following the colour scheme, figure 7 also highlights, inside rectangles, the naming convention for each cluster obtained, based on the observed patterns. It also confirms some of the suggestions arising from the previous scatter plots in figure 6. Hence, we further define the five clusters names with respect to their main observed and distinguishable patterns, as follows:

0. The 'as fast as possible' cluster: trips targeted at shortest path or even faster (e.g. faster walking/transfer/in-vehicle times than predicted), in general total time is very

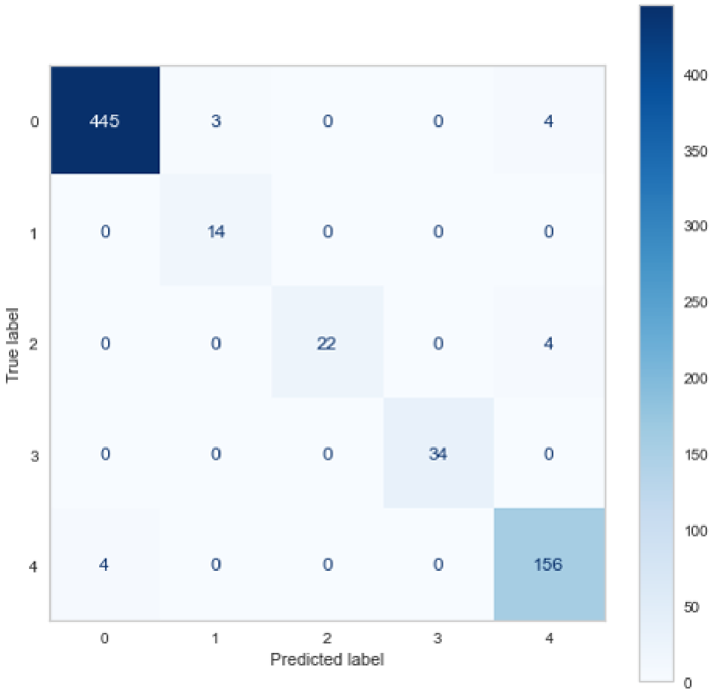
close to the shortest path counterpart;

1. The 'too early or too late for connections' cluster: trips with significantly higher transfer times than the shortest path, indicating longer waiting times for connection; in general total transit time is slightly higher than shortest path, but not too much affected;
2. The 'walking-prone' cluster: trips with significantly higher walking times than the shortest path, and lower in-vehicle and transfer detour times; as a consequence, total times also tend to be significantly higher;
3. The 'sitting-prone' cluster: trips with significantly higher in-vehicle times than the shortest path, indicating preferences for longer connections with decreased walking and transfer detour times; in general the total detour time is only slightly impacted;
4. The 'do not mind a bit of walking' cluster: trips with slightly higher walking times, but very close to shortest path in-vehicle and transfer detour times; in general this only slightly impacts total transit detour time.

Although the visualization of the results is important for gathering knowledge about the data and the differences among clusters, a proper statistical model is desired to evaluate if the patterns exhibited and inferred can be learned by a model (i.e. can a model learn to distinguish well between the different clusters based on the features provided?). By fitting a model, we can also explore how much the features contribute to each cluster.

For this purpose, the tree-boosted model with negative log-likelihood loss (as discussed in section 3) is employed. We first tune the model parameters (learning rate, max depth of trees, minimum data in each leaf and number of leaves) with a grid-based search approach and then train the model with 70% of data and test it with the remaining 30%. The multiclass confusion matrix for the test data is shown in 8, with a corresponding log-loss of 0.12.

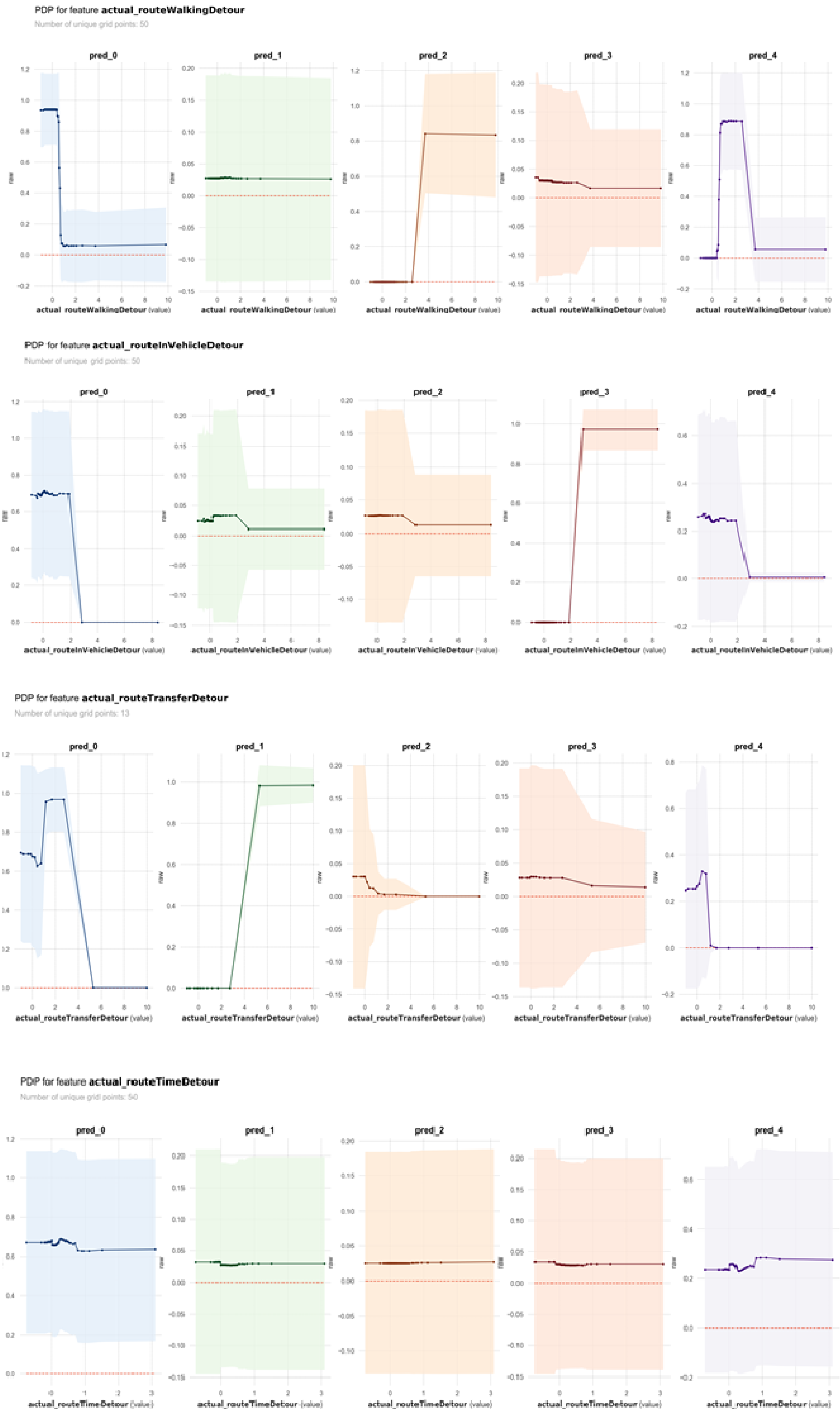
Figure 8: Tree model confusion matrix.



True label \ Predicted label	0	1	2	3	4
0	445	3	0	0	4
1	0	14	0	0	0
2	0	0	22	0	4
3	0	0	0	34	0
4	4	0	0	0	156

In terms of accuracy, the model achieves almost 0.98 on test data, showing a great ability to learn the patterns and to classify the test points into the correct cluster. In terms of the model's ability to capture the relevant features, the recall (the ratio of true positive predictions to the total actual positives) is 0.98 for cluster 0, 1.00 for clusters 1 and 3, 0.85 for cluster 2, and 0.97 for cluster 4. Similarly, precision metrics convey the model's ability in identifying positive instances, and the results are 1.00 (clusters 2 and 3), 0.99 (cluster 0), 0.95 (cluster 4) and 0.82 (cluster 1). The f1-score, a balanced measure of precision and recall, reinforces the model's robustness and the results are 1.00 (cluster 3), 0.99 (cluster 0), 0.96 (cluster 4), 0.92 (cluster 2) and 0.90 (cluster 1). In general, these results affirm the efficacy of the tree-based model in capturing the patterns in each cluster. The model's ability to distinguish between clusters with high precision and recall indicates well-defined and separable clusters, highlighting its utility in data analysis tasks. To further explore these relationships between clusters and features, figure 9 shows, for each feature and each cluster, the corresponding Partial Dependence Plot (PDP) (here with a different colouring scheme, but still ordered from 0 -left- to 4 - right).

Figure 9: Partial Dependence Plots for each feature.



PDPs are a valuable tool for interpreting the effect of individual features on the predictions of machine learning models. For multiclass classification, each plot shows information for one class/cluster (starting from class 0 on the left side to class 4 on the right side). Each feature corresponds to one full row with all classes, the order from top to bottom being: *RouteWalkingDetour, RouteInVehicleDetour, RouteTransferDetour, RouteTimeDetour*. Then, the x-axis represents the range of values for the feature of interest given the class. The y-axis represents the average predicted probability or outcome for the target class. The direction and slope of the line indicate the effect of the feature on the predicted probability of the target class; a positive slope indicates that increasing the feature value tends to increase the predicted probability of the target class, while a negative slope indicates the opposite. A flat line suggests that the feature has little to no effect on the predicted probability of the target class over its range of values.

By inspecting figure 9, it is possible to confirm many of the patterns suggested before, bearing in mind that PDPs show only marginal effects of individual features and may not capture complex interactions between features. Nevertheless, for the first feature *RouteWalkingDetour*, negative or close to 0 deviations in comparison to the shortest path are linked to very high (close to 1) probability of predicting class 0 (the 'as fast as possible' cluster). On the other hand, deviations from 0 to 2 make the likelihood of predicting cluster 4 ('do not mind a bit of walking') very high, whereas big deviations of over 2 make the prediction of cluster 2 ('walking-prone') very likely. Walking deviations do not seem to have significant effect for predictions of clusters 1 ('too early or too late for connections') or 3 ('sitting-prone').

The second row/feature is *RouteInVehicleDetour* and, by similar analysis, we see that the effects of this feature are more pronounced in clusters 0 (negative deviations to positive deviations up to 2 make it very likely to predict the 'as fast as possible' cluster) and 3 (high values of deviation starting from 2 make it very likely to predict the 'sitting-prone' cluster). Cluster 4 ('do not mind a bit of walking') is also fairly affected by detours from 0 to about 3.

The third row/feature is *RouteTransferDetour*, which has great effects on the predictions of clusters 0 ('as fast as possible') and 1 ('too early or too late for connections'), and moderate effects on cluster 4 ('do not mind a bit of walking') predictions. In particular, detour values below 4 make it very likely to predict either 0 or 4 clusters, and detour values above 4 make the prediction to be cluster 1 with probabilities very close to 1.

Lastly, we see that the variations in the feature *RouteTimeDetour* do not seem to have a great effect in the class prediction. However, this feature was included for interpretability reasons and because, in combination with the others, it increased the accuracy of the tree-based model. In general, however, cluster predictions are impacted by changes in the other features. Differences in total time happen as a consequence, but such feature does not seem to have, by itself, great prediction power to distinguish between clusters. It is interesting to analyse, however, how different route choices impact the total time, e.g. trips with higher walking time deviation will lead to a much higher total time, on average, than trips with higher in-vehicle time or transfer time deviations. Hence, from a route choice efficiency perspective, walking time is more relevant than the other features because of its potential to dramatically increase total times.

From all the analyses made before, the clusters characteristics seem to be well-defined in terms of the four variables chosen. In particular, we can highlight, for each cluster, a combination of feature values which is very likely to predict that cluster. Then, based on the resulting transit times, it is easy to visualize which features and choices are more efficient and, on the contrary, which choices are more penalizing in terms of transit times. Nevertheless, these trip attributes do not reveal whether there is an underlying user effect. As an illustrative exercise, we investigate if some users are more prone to certain behaviours when choosing their trips, i.e. we assess the proportion of each cluster among the users' trips. The results for the top-50 users (ranked based on number of trips) are plotted in figure 10.

Figure 10: Top-50 users trips segmented per cluster type.

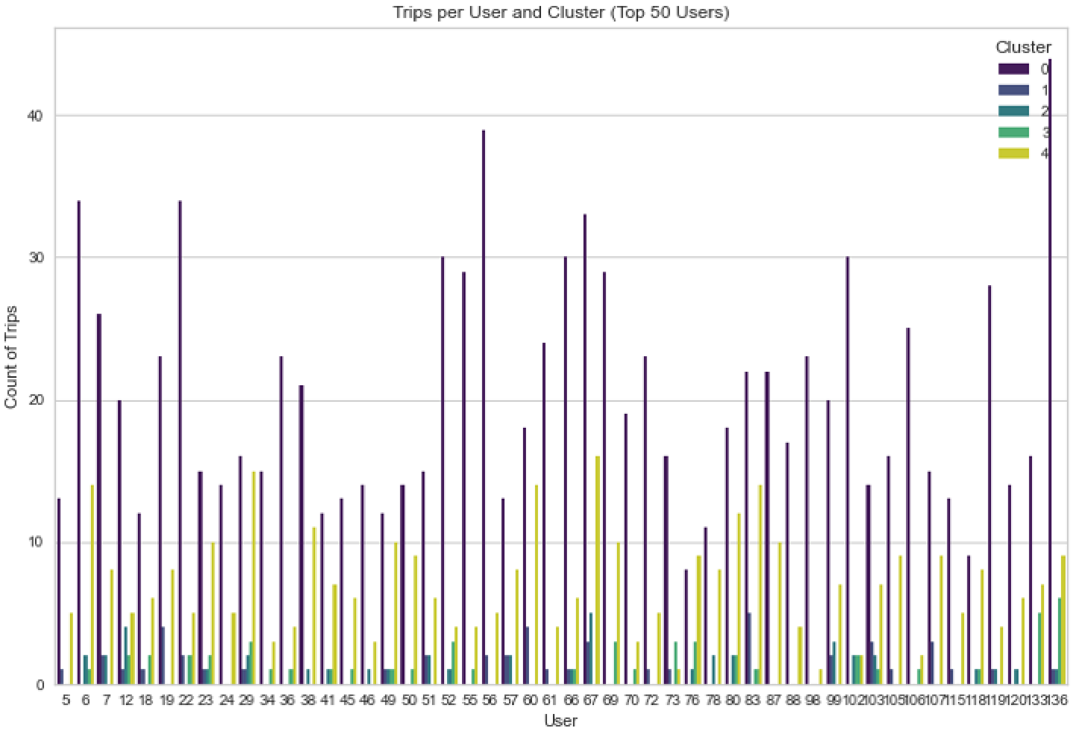


Figure 10 shows, as expected, that clusters 0 ('as fast as possible') and 4 ('do not mind a bit of walking') are the most predominant ones, encompassing most of the trips for each user. This behaviour suggests that users are aware of the shortest path trips, and aim for such trips (or trips close to them) for the most part. However, figure 10 also reveals that, in general, the user behaviour can be translated by a mix of different clusters, and not a single behaviour. Possibly other factors are impacting such decisions, such as trip purpose and other socio-demographic characteristics. This investigation would require complementing the GPS tracking data in the study with other sources of data, which is out of the scope of this paper. Nevertheless, this shows that, in a scenario where other sources of complementary data is not available, grouping trips, instead of users, is a way to measure PT route choice efficiency and some of the main attributes linked to it.

5 Conclusions and Further Research

This paper explored PT route choice efficiency in terms of metrics derived from passive GPS tracking data records. Our study employed a clustering algorithm on trajectory-based metrics to capture route choice patterns, considering both spatial and temporal dimensions. The methodology involved a three-step procedure combining machine learning methods and statistical analysis of results. Firstly, we extracted route-specific attributes from travel diaries collected by the *ETH-IVT Travel Diary* smartphone application, which provided a dataset comprising 2909 PT trips of 172 users in Zurich, Switzerland. These attributes were normalized to account for variations in trip lengths, times, and transfer patterns. Subsequently, we applied hierarchical clustering to group trips based on these normalized route attributes. Finally, we fitted a tree-based model to the clustered data and conducted statistical analysis to understand how deviations from the shortest path influence the likelihood of predicting different trip clusters according to the features analysed.

From the clustering algorithm, five clusters could be isolated based on four features of interest, namely *RouteTimeDetour, RouteTransferDetour, RouteWalkingDetour, RouteInVehicleDetour*. A tree-based model and subsequent statistical analysis in the form of PDP plots and other visualization tools confirmed the following cluster types and patterns associated with each cluster:

0. The 'as fast as possible' cluster: trips targeted at shortest path or even faster (e.g. faster walking/transfer/in-vehicle times than predicted), in general total time is very close to the shortest path counterpart;
1. The 'too early or too late for connections' cluster: trips with significantly higher transfer times than the shortest path, indicating long waiting times for connection, in general total time is slightly higher than shortest path, but not too much affected;
2. The 'walking-prone' cluster: trips with significantly higher walking times than the shortest path, and lower in-vehicle and transfer detour times; as a consequence, total times also tend to be significantly higher;
3. The 'sitting-prone' cluster: trips with significantly higher in-vehicle times than the shortest path, indicating preferences for longer connections with decreased walking and transfer detour times; in general the total detour time is only slightly impacted;
4. The 'do not mind a bit of walking' cluster: trips with slightly higher walking times, but very close to shortest path in-vehicle and transfer detour times; in general this only slightly impacts total transit detour time.

Overall, our study contributes to travel behaviour analysis by offering a framework for understanding and modelling route choice efficiency from the perspective of easy, readily available metrics from passive GPS tracking data. We focus on clusters of trips instead of clustering users. By choosing this research path, we can leverage passive GPS tracking data and offer behavioural insights even when complementary data, e.g., socio-demographics or trip purpose, are not available.

As shown in the analysis (see figure 10), users exhibit mixed behaviour towards their route choices. Although most trips will be targeted on efficiency (defined here as the shortest path), all users have some percentages of their trips that fall in some other category (e.g., longer walking times, longer in-vehicle times, etc.). Given the limitations of this study, it is not possible to know whether those choices are made rationally (on purpose) or not, and not even the factors triggering such choices. However, if we define rationality in terms of targeting the shortest path trip, we can infer that most users try to optimize their PT transit times by picking trips that do not significantly deviate from the actual shortest path trips. Most importantly, our analysis showed that the five clusters are well-defined in their feature space, so that predictions from a machine learning model are made with high accuracy. This means that the features can be used to investigate how the selected factors affect transit times. For instance, an increase in the share of cluster 1 ('too early or too late for connections') could reveal that either PT information provision is not effective, or that the planned connections are not well-adjusted. Similarly, an increase in the share of cluster 2 ('walking-prone') could indicate a behavioural shift from users (towards active-commuting) or even indicate that some connections may not be so attractive/convenient from the user standpoint (so that they still prefer to walk). For policy-makers, establishing and validating those shares and monitoring them over time can provide valuable information and metrics. On top of that, our approach is promising for aiding planners in optimizing routes and schedules to enhance efficiency and meet passenger demand.

On the basis of these findings, further research efforts should seek to explore the application of the method in a policy context, for example, identifying different clusters with varying sensitivity to policy initiatives, as suggested in Joh *et al.* (2001). Another clear direction for future research relates to complementing GPS tracking data with activity and socio-demographic information, which would provide a more comprehensive understanding of transit user behaviour and route choice dynamics, including, for instance, inference on route choice and trip patterns motivated by trip purpose and other activity-based travel features.

6 References

Carrel, A., A. Halvorsen and J. L. Walker (2013) Passengers' perception of and behavioral adaptation to unreliability in public transportation, *Transportation Research Record*, **2351** (1) 153–162.

Costa, M. A., A. D. Marra and F. Corman (2023) Public transport commuting analytics: A longitudinal study based on gps tracking and unsupervised learning, *Data Science for Transportation*, **5** (3) 15.

Dalumpines, R. and D. M. Scott (2017) Determinants of route choice behavior: A comparison of shop versus work trips using the potential path area-gateway (ppag) algorithm and path-size logit, *Journal of Transport Geography*, **59**, 59–68.

Gonzalez, M. C., C. A. Hidalgo and A.-L. Barabasi (2008) Understanding individual human mobility patterns, *nature*, **453** (7196) 779–782.

Goulet-Langlois, G., H. N. Koutsopoulos and J. Zhao (2016) Inferring patterns in the multi-week activity sequences of public transport users, *Transportation Research Part C: Emerging Technologies*, **64**, 1–16.

Goulet-Langlois, G., H. N. Koutsopoulos, Z. Zhao and J. Zhao (2017) Measuring regularity of individual travel patterns, *IEEE Transactions on Intelligent Transportation Systems*, **19** (5) 1583–1592.

Jiang, S., J. Ferreira and M. C. González (2012) Clustering daily patterns of human activities in the city, *Data Mining and Knowledge Discovery*, **25**, 478–510.

Joh, C.-H., T. Arentze and H. Timmermans (2001) Pattern recognition in complex activity travel patterns: comparison of euclidean distance, signal-processing theoretical, and multidimensional sequence alignment methods, *Transportation Research Record*, **1752** (1) 16–22.

Kim, J., J. Corcoran and M. Papamanolis (2017) Route choice stickiness of public transport passengers: Measuring habitual bus ridership behaviour using smart card data, *Transportation Research Part C: Emerging Technologies*, **83**, 146–164.

Krause, C. M. and L. Zhang (2019) Short-term travel behavior prediction with gps, land use, and point of interest data, *Transportation Research Part B: Methodological*, **123**, 349–361.

Kusakabe, T. and Y. Asakura (2014) Behavioural data mining of transit smart card data: A data fusion approach, *Transportation Research Part C: Emerging Technologies*, **46**, 179–191.

Ma, X., C. Liu, H. Wen, Y. Wang and Y.-J. Wu (2017) Understanding commuting patterns using transit smart card data, *Journal of Transport Geography*, **58**, 135–145.

Ma, X., Y.-J. Wu, Y. Wang, F. Chen and J. Liu (2013) Mining smart card data for transit riders' travel patterns, *Transportation Research Part C: Emerging Technologies*, **36**, 1–12.

Marra, A. D., H. Becker, K. W. Axhausen and F. Corman (2019) Developing a passive gps tracking system to study long-term travel behavior, *Transportation research part C: emerging technologies*, **104**, 348–368.

Marra, A. D. and F. Corman (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.

Marra, A. D., L. Sun and F. Corman (2022) The impact of covid-19 pandemic on public transport usage and route choice: Evidences from a long-term tracking study in urban area, *Transport Policy*, **116**, 258–268.

Nielsen, F. and F. Nielsen (2016) Hierarchical clustering, *Introduction to HPC with MPI for Data Science*, 195–211.

Ortega-Tong, M. A. (2013) Classification of london's public transport users using smart card data, Ph.D. Thesis, Massachusetts Institute of Technology.

Sivalingam, P., D. Asirvatham, M. Marjani, J. A. I. S. Masood, N. K. Chakravarthy, G. Veerisetty and M. T. Lestari (2024) A review of travel behavioural pattern using gps dataset: A systematic literature review, *Measurement: Sensors*, 101031.

Song, C., Z. Qu, N. Blumm and A.-L. Barabási (2010) Limits of predictability in human mobility, *Science*, **327** (5968) 1018–1021.

Thuillier, E., L. Moalic, S. Lamrous and A. Caminada (2017) Clustering weekly patterns of human mobility through mobile phone data, *IEEE Transactions on Mobile Computing*, **17** (4) 817–830.

Wang, Y., K. Qin, Y. Chen and P. Zhao (2018) Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi gps data, *ISPRS International Journal of Geo-Information*, **7** (1) 25.

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

Zhai, W., X. Bai, Z.-r. Peng and C. Gu (2019) From edit distance to augmented space-time-weighted edit distance: Detecting and clustering patterns of human activities in puget sound region, *Journal of Transport Geography*, **78**, 41–55.

Zhang, Z., Q. He and S. Zhu (2017) Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method, *Transportation Research Part C: Emerging Technologies*, **85**, 396–414.

Zheng, Y.-T., Z.-J. Zha and T.-S. Chua (2012) Mining travel patterns from geotagged photos, *ACM Transactions on Intelligent Systems and Technology (TIST)*, **3** (3) 1–18.

Zhou, Y., Q. Yuan, C. Yang and Y. Wang (2021) Who you are determines how you travel: Clustering human activity patterns with a markov-chain-based mixture model, *Travel Behaviour and Society*, **24**, 102–112.

A Appendix

Table 2: Sociodemographic and behavioural characteristics of travellers in the *ETH-IVT Travel Diary* survey

Variable	Category	Counts
Main Occupation	Employed / self-employed	107
	Student	41
	Student and Employed / self-employed	11
	Other	11
Average time spent from home to work [min]	$[0, 15]$	51
	$(15, 30]$	72
	$(30, 45]$	25
	> 45	21
Work/School days	Weekdays Only	145
	Everyday	10
	Weekdays and Saturday	7
Driver's license ownership	Yes	119
	No	50
PT subscription	Yes	127
	No	38
Frequency of use (PT and Private Modes)	PT = almost daily; bike/car/private = 1-3 days per week	58
	PT = almost daily; bike/car/private = 1-3 days per month	26
	PT = almost daily; bike/car/private = almost daily	20
	PT = 1-3 days per week; bike/car/private = almost daily	14
	PT = almost daily; bike/car/private = rarely to never	13
	PT = 1-3 days per week; bike/car/private = 1-3 days per week	12
	PT = 1-3 days per month; bike/car/private = almost daily	8
	PT = 1-3 days per month; bike/car/private = 1-3 days per month	5
	PT = 1-3 days per week; bike/car/private = 1-3 days per month	4
	PT = 1-3 days per month; bike/car/private = rarely to never	3
	PT = 1-3 days per month; bike/car/private = 1-3 days per week	2
	PT = 1-3 days per week; bike/car/private = rarely to never	2
	PT = rarely to never; bike/car/private = almost daily	1
	PT = rarely to never; bike/car/private = 1-3 days per week	1
	PT = rarely to never; bike/car/private = rarely to never	1