Treemob: Expressive Mobility Data Representation Through Tree-Based Structures

Marta Fioravanti[1], Eleonora Cappuccio[1,2,3(✉)], Salvatore Rinzivillo[2], and Riccardo Guidotti[1,2]

[1] University of Pisa, Pisa, Italy
m.fioravanti6@studenti.unipi.it, eleonora.cappuccio@phd.unipi.it, riccardo.guidotti@unipi.it
[2] ISTI-CNR, Pisa, Italy
{eleonora.cappuccio,salvatore.rinzivillo,riccardo.guidotti}@isti.cnr.it
[3] Università degli studi di Bari Aldo Moro, Bari, Italy

Abstract. This paper explores expressive representations of personal mobility data, focusing on both informational and user-centered design aspects. The goal is to enable users to access, understand, and gain awareness of their mobility behaviours, assuming that expressiveness is not inherently linked to system complexity. We propose a novel methodology for representing and analyzing mobility data using tree-shaped structures. Additionally, we introduce Treemob, a suite of Python-based tools designed to facilitate mobility analysis. The experiments conducted provide a foundation for further research, offering a flexible framework for exploiting different issues and contexts.

Keywords: Mobility Data Analysis · Individual Mobility Network · Spatio-temporal data visualization

1 Introduction

Nowadays, vast amounts of personal data are generated, including spatio-temporal information about individual mobility. These data offer valuable insights, but their raw form requires organization for effective analysis, especially in the context of Personal Data Analytics [13]. Technologies such as GPS and GSM data enable the study of mobility patterns, contributing to urban planning, traffic analysis and sustainable transportation solutions. For example, Andrienko et al. [2] used visual analytics to identify significant geographic locations from trajectory data, while Larcom et al. [12] observed behavioural changes in response to unexpected travel disruptions. These studies demonstrate the potential of mobility data to uncover behavioural patterns at both the individual and collective levels. Recent work has shifted from global mobility analysis to individual models, which reveal detailed, systematic behaviours.

A. Kocian et al. (Eds.): INTSYS 2024, LNICST 608, pp. 46–63, 2025.
https://doi.org/10.1007/978-3-031-86370-7_4

Studies like Pappalardo et al. [14] identified distinct traveller types-explorers and returners-while Trasarti et al. [17] explored practical applications such as carpooling through personalized mobility analysis. Finally, Landi et al. [10,11] designed novel features extraction methodologies to capture interpretable and highly discriminative mobility patterns to boost trajectory classification models. Individual models not only enhance user-centered data applications but also raise ethical concerns about privacy, necessitating frameworks that allow users to control the sharing of their data. However, individual models requires a suitable level of abstraction to let the user understand their data without overwhelming them with excessive cognitive load.

This research introduces a novel method for summarizing and visualize personal mobility data using hierarchical structures, to bridge the complexity of network-based methods [5,8,20]. We complement this analysis with a visual analytics tool and a Python module that facilitates tree-based mobility data analysis. The objective is to apply unsupervised learning to identify mobility patterns across different cities.

2 Related Work

Mobility data analysis can be performed from collective or personal perspectives. In the collective model, the trajectories of multiple users are aggregated to analyze global patterns, such as traffic flows or congested urban areas. Collective mobility data has been instrumental in fields like urban planning, where understanding the movement of large populations over time helps inform decisions about infrastructure improvements [18]. However, as noted by Trasarti et al. [17], global patterns often fail to capture the diversity of individual behaviours, leading to oversimplifications in complex systems like human mobility.

To address this limitation, personal mobility models have been proposed. One example is the mobility profile extraction pipeline introduced by [17], which identifies user habits by clustering trajectories that are temporally and spatially similar. This method aggregates frequent trips into a set of prototypes, effectively creating a mobility profile that reflects the habitual movement patterns of the user. Another significant approach is the Individual Mobility Network (IMN) [7], which organizes user behaviour into a directed graph structure. In the IMN, vertices represent regular and irregular locations, and edges denote the movements between them. This structure provides a nuanced view of personal travel patterns, distinguishing habitual movements from outlier trips. Combined models, which blend global and personal perspectives, have also emerged to exploit the strengths of both approaches by creating hybrid models.

Visual analytics tools play a crucial role in the exploration and interpretation of mobility data. Andrienko et al. [3] proposed a comprehensive visual analytics methodology to analyze both individual and collective movement behaviors. In this approach, movement data is not only visualized geographically but also in a space-time cube, where the third dimension represents the temporal component. This multi-dimensional representation allows users to observe trajectories over

time, identifying patterns that might not be visible in a purely spatial view. Moreover, the authors introduce a transformation that maps movements into an abstract space, where directions and clustering of similar trajectories can be analyzed, providing deeper insights into group and individual behaviors. However, one drawback of this tool is its complexity, which can make it challenging to apply outside of research contexts or for communicating results to non-expert audiences.

In another work, Andrienko et al. [1] addressed privacy concerns in mobility data visualization by proposing an abstract representation that hides sensitive user information. The authors developed a pair of two-dimensional histograms that display probabilities for a user's presence at different locations over time, without revealing precise spatial details. The visual simplicity of this method makes it an effective qualitative tool for summarizing mobility patterns while preserving user privacy. However, the reliance on square area sizes to represent probabilities can make it difficult to compare values accurately, especially for less technical users.

A different approach to visualizing mobility data was introduced by Douieb[1], who focused on representing football passes within a play area. This flow visualization emphasizes spatial patterns by displaying movement paths without focusing on individual players. While this tool is designed for a specialized application, it illustrates the potential for visualizing high-density movement data in other contexts, such as public transportation networks or pedestrian flows in urban environments.

3 Preliminaries and Background

This section reviews the key approaches and methodologies used to represent, analyze, and visualize mobility data, focusing on both personal and collective perspectives, as well as the various techniques employed to simplify and interpret complex trajectory datasets.

3.1 Mobility Data Representation and Analysis

The fundamental element in any mobility dataset is the trajectory, defined as an ordered sequence of points traversed by a user over time. Formally, a trajectory t_s can be defined as a series of spatiotemporal points $t_s = \langle p_1, ..., p_n \rangle$, each of them in the form of a triple $p_i = (lon_i, lat_i, t_i)$ where lon_i and lat_i are respectively the longitude and the latitude of the location traversed, while t_i is the timestamp of the event. A trajectory is chronologically ordered, so $\forall 1 \leq i \leq n : ts_i < ts_{i+1}$; so, the i^{th} point of the trajectory ts is identified as $ts[i]$, while its spatiotemporal coordinates are $ts[i].lon, ts[i].lat$, and $ts[i].t$. Therefore, a mobility dataset can be seen as a set of trajectories $T = \langle t_i, ..., t_m \rangle$.

[1] https://observablehq.com/@karimdouieb/all-the-passes.

3.2 Tree and Graph-Based Representation

Graphs and trees are fundamental structures for representing and analyzing mobility data, since they are capable of capturing the relationships of different dimensions of human mobility. For example, the origin-destination matrix of a user can be represented with a **mobility graph**, i.e. a directed graph $G = (V, E)$ where V is the set of nodes representing locations, and E is the set of edges denoting movement paths between those locations. Each edge $e = (v_a, v_b)$ corresponds to a trajectory from location v_a to v_b. The edges may be enriched with additional information, such as weights, to reflect the frequency or importance of these paths.

Another common structure in mobility data analysis is the **prefix tree** typically used to compress large sets of sequences. A prefix tree is a variation of a hash tree and can be defined as $PT = (V, E, root)$ where V represents nodes (locations), E represents edges (paths), and the $root$ is a virtual node not associated with any location. The defining feature of a prefix tree is its ability to group all sequences sharing a common prefix into a single branch. This allows for efficient storage and retrieval of sequences, making it ideal for summarizing user trajectories. In [19], Zhao et al. combined prefix trees with differential privacy techniques to create trajectory representations that maintain the structural integrity of mobility data while ensuring user privacy by adding noise to sensitive information.

Spanning graphs aim to simplify the representation of trajectories by preserving essential paths while reducing complexity. A spanning tree is a subgraph of a weighted graph that retains the shortest paths between nodes with minimal error. In the context of mobility data, spanning trees can be used to focus on frequent connections between locations rather than rare or outlier movements. When applied to personal mobility, maximum spanning trees are particularly useful, as they emphasize the most commonly traversed paths, which are often more relevant for understanding user habits.

4 Problem Formulation

In this section, we formulate the problem of transforming a user's mobility data, represented as a set of trajectories $M_u = \{t_1, \ldots, t_n\}$, into two distinct data structures: a tree $T = \{N, E\}$ and a vector $V(T)$. These transformations allow us to encode both trajectory information and structural characteristics of the mobility data in a compact form. Below, we detail the two approaches: tree-based and vector-based representations.

4.1 Tree-Based Representation

The first step in our formulation is to convert a set of user trajectories into a tree structure $T = \{N, E\}$ where $N = \{n_1, ..., n_p\}$ is a set of nodes corresponding to locations visited by the user, and $E = \{e_1, ..., e_m\}$ is a set of directed edges between these nodes.

Prefix Tree. The prefix tree is an intuitive representation of trajectories, treating each trajectory as a sequence of spatial points. Each point, simplified to contain only spatial data, is represented as a node, and the edges reflect transitions between these points in chronological order. Of course, this approach is a trade-off between completeness of information and accessibility. This level of simplification of real mobility masks certain details to provide an overview and let the analyst focus on specific details. In fact, a prefix tree is a flexible structure that can be dynamically explored to highlight different perspectives.

Specifically, we assign each location $g_i = (lon_i, lat_i)$ a unique identifier $j \in J$, where each $j \in J$ is a unique identifier of a spatial grid, for example a regular grid, a voronoi tessellation or a geohash code. Thus, a trajectory point previously defined as $p_i = (lon_i, lat_i, t_i)$ is now simplified as $p_i = j_i$. The resulting tree will be composed by the nodes $N = <n_1, ..., n_n>$, the labels $L = <l_1, ..., l_n>$, assigned to each node and the edges $E = <e_1, ..., e_m>$. The nodes are artificial structures associated with a unique identifier, and contain the information of the represented location; the same label value can be present in different nodes in various positions in the tree. This approach yields a compact yet powerful representation of the user's movements.

Spanning Tree. The prefix tree requires a set of sequences to be scanned during the growing of the tree. For mobility data a natural choice may be the temporal order of the spatio-temporal observations of each trajectory. However, this naive approach may lead to configurations where the first encountered positions may determine the general shape of the whole tree. For example, if the logging of positions starts from a low relevant point, this choice may raise the corresponding position very in high in the tree. Since we have the mobility graph data structure, we propose the use of a spannign tree algorithm to derive the prefix tree given a mobility graph. In particular we adopted the Kruskal algorithm [9], to find a spanning forest over a mobility graph.

Since we are interested in the most frequent paths, we use as a weighting function the inverse of their frequency to keep the algorithm a minimum spanning tree. The resulting representation synthesises the user's mobility, focusing on the connections between each location. This structures preserves the uniqueness of each location/node, since no repetions are possibile. This creates a more condensed data structure and it also keep into account the relation of each location/node with the mobility graph.

4.2 Vector-Based Representation

The second transformation focuses on representing the mobility data as a vector. The vector representation allows us to encode the structure and patterns of mobility in a form suitable for numerical analysis.

Tree-to-Vector Transformation. To convert the tree structure T into a vector $V(T)$, we employ techniques inspired by information retrieval, specifically

the TF-IDF (Term Frequency-Inverse Document Frequency) approach. In this analogy, locations in the mobility data correspond to "terms" and users to "documents". The frequency of visits to locations is treated as the term frequency (TF), and the inverse document frequency (IDF) measures how uniquely a location characterizes the user compared to others. The resulting vector $V(T)$ highlights distinctive locations and paths that differentiate a user's mobility from that of others. This vector representation maintains the focus on the shape of the user's mobility, abstracting away from geographical coordinates while still preserving critical information about movement patterns.

Feature Extraction for Mobility Analysis. Additionally, we propose an alternative method of vector representation by extracting features from the tree structures themselves. By summarizing the properties of the mobility trees-such as the number of branches, depth, or the frequency of node occurrences-we generate feature vectors that can be used in unsupervised learning or other analytical tasks. These vectors enable us to quantify and compare different users' mobility patterns without accessing the raw trajectory data.

In summary, both the tree and vector representations provide powerful tools for analyzing personal mobility data. The tree captures structural and sequential information, while the vector form enables further analysis through numerical methods. Together, these approaches offer a comprehensive solution for representing and studying mobility patterns.

5 Methodology

This section outlines theoretical and technical methods for transforming trees and vectors into semantically meaningful representations, focusing on the role of the root in trees and the relationships encoded in both trees and vectors.

5.1 Tree-Based Methods

A trajectory notation was introduced that adds artificial points ('^' and '$', representing respectively the start and the end of a trajectory) to preserve trajectory direction when organized in prefix trees. For prefix trees, three rotation strategies centred around a root location (the most frequent in the dataset) were evaluated. The chosen structure separates trajectories that touch the root from those that do not, facilitating a clear analysis of movement around the root. For spanning trees, undirected graphs were transformed into rooted trees to focus on spatial relationships around locations rather than movement sequences.

5.2 Vector-Based Methods

Two vectorization approaches were developed to represent mobility data. The first, a TF-IDF-inspired method, weighs locations based on their frequency and

proximity to the root, adding semantic meaning to the analysis of mobility. The second, a feature-driven method, extracts nine features from mobility trees (e.g., depth, unique location ratios) to create vectors that allow for clustering and comparison. This vector-based approach provides a more abstract and feature-driven perspective on user mobility, complementing the tree-based analysis.

5.3 Distance Functions

Distance functions were applied to compare trees and vectors derived from mobility datasets, enabling the identification of clusters of similar users. Unlike vectors, tree-shaped structures require specialized distance functions. The edit distance is a dynamic programming method that calculates the minimum number of operations (insertion, deletion, relabeling of nodes) required to transform one tree into another. A variation of this method, which compares branches as unordered sets rather than sequences, offers polynomial-time computation and improves on the standard version's exponential complexity. To further enhance the analysis, an ordering mechanism was introduced where heavier branches, based on frequency, are positioned on the left. This allows for easier comparison of tree structures. Additionally, location labels were removed from the nodes, focusing comparisons on user movement patterns rather than specific locations visited. For vectors, traditional distance measures like Euclidean, Manhattan, and Minkowski were employed, given their suitability for vector-based representations.

5.4 Visual Representation of Mobility Tree

Data visualization is essential for both exploring data and effectively communicating research findings. To achieve this, a visual system was designed using the D3.js library, a widely used tool for interactive data visualization. This system allows for the exploration and representation of mobility trees derived from user data, making the complex structure of personal mobility accessible. By implementing the tool in D3.js, a high level of customization was achieved, enabling the creation of interactive modules that visualize various aspects of user mobility in an intuitive and engaging way.

The visualizations are presented in a hierarchical manner, consistent with Shneiderman's Visualization Mantra [16], where more general and important components are immediately visible, while more detailed information is accessible on demand. The primary output is a node-link layout, which mirrors the underlying data structure. In this layout, nodes represent locations, and links depict the paths between them. Nodes are scaled based on how frequently each location is visited in the user's dataset. To enhance interactivity, hovering over a node highlights all corresponding locations across the tree, allowing the discovery of movement patterns. A variation of this tree layout is also provided, where the lengths of links represent the actual distances travelled, offering a clearer view of travel patterns in terms of spatial distance (see Fig. 1).

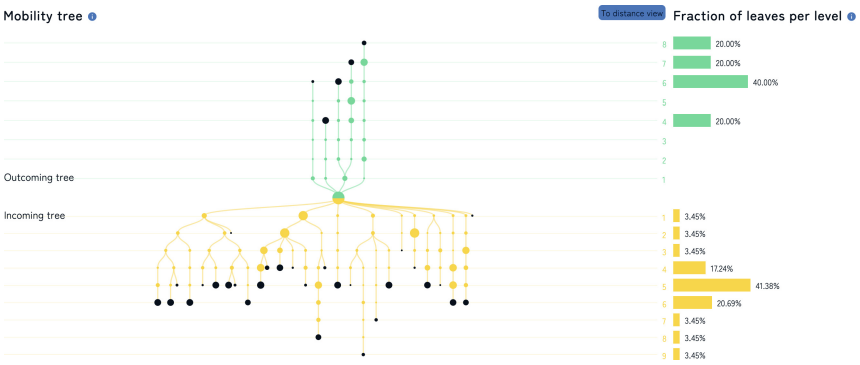

Fig. 1. The mobility prefix tree in the classical shape, accompanied by the bar chart, represents the number of terminal locations for each level.

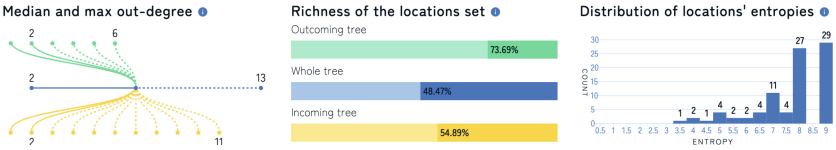

Fig. 2. From left to right: The median and maximum out-degree visualisation; the progress bar representing the ratio between locations and total nodes; the distribution of the locations' entropies.

To complement the tree structure, additional components offer further insights into user mobility. For example, a leaf exploration tool displays the percentage of leaves at different depths, reflecting how far users tend to travel. Barcharts, positioned alongside the tree, represent the richness of location diversity, indicating how often users visit new places versus recurring locations. Out-degree diagrams (Fig. 2) further break down user movements by showing the median and maximum number of connections from each node, revealing the complexity of travel routes around central locations. Another component of the visualization is a progress bar representing the richness of the location set, meaning the proportion between the number of different locations encountered and the number of all the traversed points. The visualization system also includes a histogram displaying the local entropy of locations, which quantifies how much information or variability a location contributes to the user's overall mobility. This is particularly useful for identifying high-entropy locations that play a significant role in the user's travel patterns. The histogram interacts with the main tree visualization, highlighting relevant entropy values when a node is hovered.

We intentionally excluded the use of maps to abstract the geographic layer and protecting user privacy, aligning with ethical considerations in mobility data analysis. The visualization still effectively captures movement patterns and behaviours without relying on geographic context. This choice not only protects

privacy but also demonstrates how mobility data can be meaningfully represented through innovative visual strategies. By focusing on user habits and travel patterns, the system offers a comprehensive, anonymized view of personal mobility, providing researchers with a powerful tool for analyzing complex movement data while adhering to privacy standards.

6 Experiments

Mobility Trees offer the feature of summarizing and representing in a concise and ordered way the mobility of an individual. This summarization may enable the identification of groups of individuals with similar mobility patterns and the comparison of different cities to identify distinct movement behaviors. In this section we show the results of the experiments conducted on a dataset of GPS-tracked vehicle movements within two main cities in Tuscany, Italy: Pisa and Florence. For comparison, we used both the Tree-based and Vector-based encodings to identify the most suitable clustering algorithm for the mobility data.

6.1 Preprocessing

To address the extensive number of locations and trajectories typically associated with each user, several data-cleaning techniques were implemented to reduce both computational time and noise that could impact the analysis.

Selection of Locations. For this analysis, two cities, Pisa and Florence, were selected as prototypes. These cities vary in size and user population, providing a solid foundation for detecting potential differences in mobility patterns. From the GPS dataset, a square perimeter surrounding each city was defined, encompassing adjacent areas to account for users residing outside the exact city boundaries.

User and Trajectory Pruning. The preprocessing of each user's dataset involved multiple steps to eliminate outlier trajectories and remove profiles lacking sufficient data for comparison. Initially, locations within the defined perimeter were selected from the seed file. Users whose mobility was predominantly (at least 65%) confined to the city were retained. This percentage was calculated by comparing the size of the user's trajectories within the perimeter to the original dataset size. This approach ensured that users with limited mobility to a few locations were not penalized, excluding infrequent travels covering extensive locations within the city.

Subsequently, user-specific data cleaning was performed to remove outlier trajectories based on length. The distribution of traversed points for each trajectory was analyzed, and outliers were identified and discarded using the interquartile range (IQR) method. Length in terms of traversed locations was prioritized over

distance travelled in meters, as data tessellation indicated an average distance of approximately 200 m between adjacent points, providing a reasonable approximation.

Global outlier removal followed, extracting the number of trajectories and median length for each user. After converting these values to a logarithmic scale, the IQR test was reapplied to both distributions, resulting in the exclusion of further outliers.

For the Pisa dataset, the initial count of users was 1,895, which was reduced to 1,636 after the cleaning process. The initial distribution of trajectories per user revealed a mean of 588.79, a median of 387, and a standard deviation of 705.1; after cleaning, these values adjusted to 637.1, 501, and 548.7, respectively. The trajectory length distribution initially exhibited a mean of 29.15, a median of 18, and a standard deviation of 51.02, which changed to 23.73, 17, and 24.5 post-cleaning.

In contrast, the Florence dataset began with 7,155 users, decreasing to 5,762 after cleaning. The original distribution of trajectories per user had a mean of 487.47, a median of 255, and a standard deviation of 841.48, which were adjusted to 553, 386, and 532.697, respectively, after preprocessing. The lengths of trajectories initially had a mean of 25.67, a median of 14, and a standard deviation of 49.42, which changed to 20.468, 14, and 22.795, respectively.

6.2 Tree Extraction

Upon completion of the cleaning process, user-specific mobility trees were generated. Given the volume and complexity of the data, the Ramer-Douglas-Peucker (RDP) [15] algorithm was employed to simplify trajectories. This algorithm approximates paths by removing points that do not significantly differ from preceding and following points in terms of direction. An epsilon value of 0.001 was determined to provide a satisfactory balance between approximation and information retention, as alternative values of 0.005 and 0.0005 yielded similar conceptual results.

From the cleaned data, mobility prefix trees for each user were generated. Due to the computationally intensive nature of clustering algorithms, which require iterative distance calculations between points, a distance matrix between trees was pre-computed using the unordered tree edit distance.

To compare the mobility patterns across different cities, a z-test was conducted to determine whether statistically significant differences existed in the distributions of vector attributes. The attributes examined included the median and maximum depths of both the incoming and outgoing subtrees, the median and maximum out-degrees of the nodes, and the ratio of unique locations to the total number of nodes. The resulting tests are showed in Table 1.

The mobility dataset was preprocessed to remove sensitive information and to generalize movements on a spatial grid determined by a voronoi tessellation. Each movement represented by a sequence of raw GPS points was transformed into a sequence of grid cells. This generalitazion was necessary to protect user privacy and to allow for a more abstract representation of the data. The dataset

Table 1. Pisa's and Florence's distributions

Pisa's and Florence's distributions				
Attribute	Test statistic	P-value	Mean Pisa	Mean Florence
Median depth incoming tree	1.606	0.108	8.0767	7.898
Median depth outcoming tree	−2.531	0.011	1.559	1.699
Max depth incoming tree	0.94	0.347	24.493	24.181
Max depth outcoming tree	1.937	0.053	20.287	19.576
Locations/nodes ratio	−3.946	0.0007	0.15	0.164
Median incoming tree out-deg	0.533	0.594	2	2
Median outcoming tree out-deg	0.571	0.568	1.955	1.952
Max incoming tree out-deg	0.571	0	14.392	12.666
Max outcoming tree out-deg	4.854	0.00001	9.833	9.148

after this phase consists of three primary tables: Seed, which associates grid cells location IDs with their geographic coordinates; *Trajlinks*, which records trajectory information across user IDs; *Trajstats* which store movement attribute distribution, including travel date, duration, and distance.

In the analysis phase, we focused on the prefix tree representation to showcase its potential for summarizing and comparing mobility patterns. By concentrating on this less-explored structure, the experiment aimed to evaluate the feasibility of using tree-shaped data for analysis and to identify potential challenges analysts might encounter when working with such representations. Moreover, the tree analogy provides significant opportunities for effective scientific communication and data storytelling.

6.3 Preliminary Test

Before the main analysis, several feasibility tests were conducted to assess the computational time required for each method. A batch of users from the analysis dataset, particularly those with more trajectories, was selected to have an estimate of the computational cost, allowing for the selection of feasible techniques.

Initial tests showed that the basic operations were computationally efficient. For example, the heaviest tree, containing 681 trajectories, 5428 locations, and an average trajectory length of 84.9 points, was generated in less than half a second. A rotation operation on the tree has the objective of selecting a different root node associated with a different location. Even this operation took no longer than 1.1 s.

Deeper tests on other operations were also performed. Four prefix trees were built for each user based on random samples of 25%, 50%, 75%, and 100% of their trajectories. These tests aimed to evaluate how the data volume impacted performance. The standard tree edit distance was found unsuitable for large-scale real-world datasets, as the tree ordering algorithm took between 103.4 and

Table 2. Completion time of Tree Ordering using different methods: raw ordering, ordered edit distance, unordered edit distance. Completion time expressed in seconds

Percentage of Data	Ordering	Ordered Tree ED	Unrdered Tree ED
25%	103.4099	10.4926	4.3097
50%	204.7914	43.6611	16.0067
75%	256.842	83.3877	26.2233
100%	559.8551	353.147	128.5133

560 s, while the distance function ranged from 10.5 to 353 s. Although optimizing the ordering algorithm might have been possible, reducing the complexity of the distance function remained a challenge (see Table 2).

In contrast, the unordered tree edit distance initially ranged from 4.3 to 128.5 s. However, after optimizing the adjacency matrix structure, the mean computation time on actual data dropped to 0.4 s. This improvement was also supported by data cleaning and selection steps.

Given the large number of techniques hypothesized in the preliminary phase, a choice was made between covering all methods to provide broad but superficial results or focusing on a specific set for a deeper investigation, even if tested methods are incomplete. The decision was made to explore the limits of the tree-shaped representation, with the adopted methodology easily extendable to other scenarios.

Consequently, the prefix tree structure was selected over the spanning tree, and a feature-based vector representation was chosen instead of a TF-IDF approach. The prefix tree architecture was considered more interpretable and suitable for narrative purposes, while the feature-based vector representation provided a complementary perspective within the same analytical framework.

6.4 Clustering Algorithms

The initial choice for clustering algorithms centred on the k-medoids method, primarily due to the need for representative trees for each cluster and the unsuitability of artificial points as centroids for tree data. A grid search was conducted to determine the optimal value of k, measuring the silhouette score for values ranging from 2 to 30 across 30 random initializations of medoids. Additional consideration was given to the sample size within each cluster and the silhouette scores of individual clusters to gain a nuanced understanding of the clustering dynamics. Alternative algorithms, such as DBSCAN [6] and OPTICS [4] for density-based clustering, as well as various hierarchical clustering methods including single, complete, and centroid linkage, were also explored.

Clustering Trees. For both Pisa and Florence, the k-medoids clustering, utilizing the unordered tree edit distance, consistently yielded a predominant cluster

characterized by a high average silhouette score (0.71 for Pisa and 0.62 for Florence), alongside sparse and fragmented clusters with negative silhouette scores, regardless of the value of k. The most favourable results were generally achieved with $k = 4$ This trend persisted across different cleaning techniques and varying RDP epsilon values, leading to the retention of the configuration that maximized silhouette scores, as previously detailed in the preprocessing section. The primary cluster for Pisa encompassed an average of 1,481.5 users, while the cluster for Florence averaged 2,552 users, indicating a significant number of individuals with similar movement patterns. To validate these findings, DBSCAN and OPTICS were also employed, revealing similar clustering phenomena. The epsilon parameter for DBSCAN was selected based on a plot of distances to the $i - th$ neighbour, ultimately retaining the distance to the 4th neighbour as recommended by the algorithm's authors; the epsilon values were set at 3000 for Pisa and 2000 for Florence. DBSCAN successfully aggregated 96% of the Pisa samples into a single cluster, while 97% of Florence samples fell into the same category. Similar clustering results were obtained with OPTICS, and all hierarchical clustering techniques tested indicated the emergence of a dominant cluster comprising a substantial subset of the dataset.

Clustering Vectors. Subsequently, the analysis transitioned to clustering the vectorial abstractions of the prefix trees, aimed at uncovering broader patterns in user mobility while disregarding geographical aspects. The k-means algorithm was employed in conjunction with various feature projection techniques. A grid search was performed to identify the optimal number of clusters (k) within a range of 2 to 6, computing the average silhouette for each configuration. This process was repeated for standardized vectors and those projected through independent component analysis (ICA), principal component analysis (PCA) with 2 to 4 features, and multidimensional scaling. For both cities, the best outcomes were observed with a 2-means clustering on the 2-dimensional PCA projection, yielding silhouette scores of 0.45 for Pisa and 0.49 for Florence. However, no meaningful clusters were identified across the configurations tested; even the highest silhouette scores represented divisions within a larger, dense agglomerate of points corresponding to the majority of users. Figure 3 visualizes the results of the clustering for both cities.

Consequently, the second-best result – a 3-means clustering on the 2-dimensional PCA projection – was selected to investigate potential differences within this space and provide three user prototypes, despite the overarching presence of a large and dense group.

The trees closest to each cluster's centroid were analyzed for both cities, and their distinct characteristics are summarized as prototypes P0, P1, and P2 for Pisa, and F0, F1, and F2 for Florence. In Pisa, P0 exhibited a mixture of branches containing both frequently visited and infrequently visited locations, with notable variability in node sizes. P1, on the other hand, was dominated by larger nodes, reflecting a concentration of frequently visited locations with lower location richness and consistent visitation patterns. This regularity was further

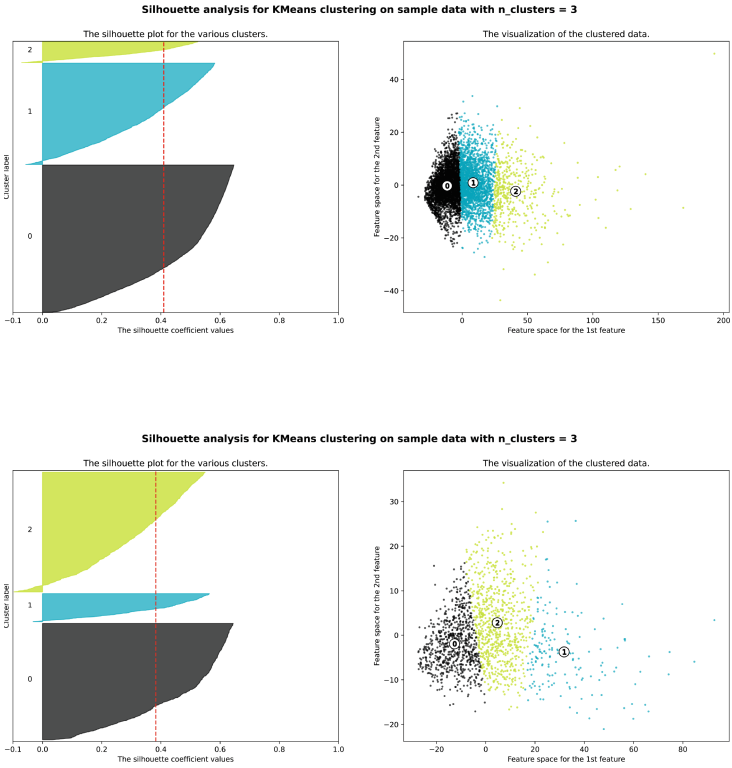

Fig. 3. The results of the 3-means on the vectors projected in two dimensions through PCA.

supported by the limited variation in node sizes, corresponding to higher entropy values. In contrast, P2 consisted predominantly of smaller nodes, indicating a broader and less predictable set of locations, with greater location richness and lower frequency variability. Additionally, P0 featured a prominent branch likely representing the user's regular commute, highlighting a mixed visitation pattern that combined both frequent and rare locations.

In terms of node out-degree, the three prototypes in Pisa exhibited similar characteristics. The visualization mapped against cumulative distances traveled in each path revealed no substantial differences regarding maximum distances traveled in individual trips. All prototypes showed a similar distribution of leaf distances, with the majority of travels concluding within the first seven kilometers. As the distance from the root increased, the frequency of observed travels diminished.

For Florence, the differences in location frequencies were less pronounced, with F0 displaying more variance and a slightly higher maximum entropy value. The prototypes exhibited comparability in terms of location richness and out-degree. The quantity of possible prefixes appeared more discriminative in Flo-

rence; F1 exhibited a notably reduced tree, while F0's root had a vast array of successors, indicating significant prefix variety. F2 was larger relative to F1 but smaller compared to F0. An analysis of travel distances revealed greater diversification in Florence users than in Pisa, with the distribution of leaves relative to distance from the root showing less gradation and identifying specific distance ranges frequently traversed.

Another qualitative assessment involved mapping each user's root by cluster within each city's geographical context to explore potential spatial correlations. However, no clear geographical distinction emerged between groups in either Pisa or Florence.

6.5 Results

Further analysis was conducted to compare the characteristics of trees from the two cities. This involved examining the features used to encode the trees into vectors for the previous step and applying a z-test for each dimension after standardization. While the test indicated statistically significant differences in means for certain attributes, the reliability of these results was questionable due to skewed distributions. Table 1 illustrates that differences in means were relatively minor, with the most notable variation occurring in the out-degree of outgoing trees, where differences of approximately two units and 0.15 in median depth were observed. Yet, these differences were not substantial. Contrary to initial expectations, distinct categories of users were not identified within the tree-shaped data or the derived vector representations. However, by examining the most central samples in each cluster for both cities, several distinguishing characteristics emerged. This observation suggests a continuous data space, implying that there are intermediate user types. This hypothesis is reinforced by consistent results across varying preprocessing methods, algorithms, and data structures. Notably, the distinguishing factor appears to be the regularity of users' paths: among the observed users, at least one exhibited equi-probable node usage across branches, while others displayed frequent and rare locations. This characteristic, particularly evident in Pisa, aligns with the "returners" and "explorers" framework proposed by Pappalardo et al. [14]. Additionally, the lack of correlation between users' root positions and their cluster memberships further supports the notion that clustering is not biased by the locations traversed.

7 Conclusion and Future Work

This research explored the mapping of personal mobility data into tree structures, specifically utilizing a variant of prefix trees rotated around the most frequent location. The tree unordered edit distance served as the distance function, enabling effective comparisons of trees irrespective of branch order. Additionally, a vectorial description of the trees was developed to capture their structural

characteristics. The analysis focused on the cities of Pisa and Florence, where unsupervised learning was employed to identify distinct user types based on both tree and vector representations. The results consistently indicated the presence of a dominant dense group with minimal noise across all tests. We recognize that the focus on two cities with distinct population sizes and user demographics may limit the generalizability of findings. While these cities offer an initial foundation for exploring mobility patterns, some characteristics-such as infrastructure and population density, may not fully represent the range of mobility dynamics found in diverse urban forms or regions. Consequently, further research will be needed to assess the applicability of tree representations in cities with more complex transportation systems and varied commuting behaviors.

Future research directions include leveraging visualization tools and establishing semantically diverse locations as roots to gain deeper insights into human mobility patterns. This approach could facilitate comparisons of trees generated from distinct data sets belonging to the same individual, such as weekday versus weekend mobility or seasonal variations. Furthermore, examining mobility patterns across different cities in various countries or from alternative modes of transport may yield valuable insights. Another avenue for exploration involves clustering movements within a single tree to uncover recurrent patterns in user behavior. To enhance accessibility and user-friendliness, future work should also consider user-centred design improvements, including user and heuristic usability testing, to ensure that end-users can easily interpret and interact with tree-based mobility data representations. Overall, this work underscores the potential of representing mobility data as trees, reinforcing the findings of [14], which suggest a dichotomy between returners and explorers.

Acknowledgements. This work is partially supported by the EU NextGenerationEU programme under the funding schemes PNRR-PE-AI FAIR (Future Artificial Intelligence Research) - Spoke 1- Partnership Extended PE00000013, PNRR-SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics - Prot. IR0000013, H2020-INFRAIA-2019-1: Res. Infr. G.A. 871042 *SoBigData++*.

References

1. Andrienko, G., Andrienko, N.: Privacy issues in geospatial visual analytics. In: Advances in Location-Based Services: 8th International Symposium on Location-Based Services, Vienna 2011, pp. 239–246. Springer (2012)
2. Andrienko, G.L., Andrienko, N.V., Hurter, C., Rinzivillo, S., Wrobel, S.: From movement tracks through events to places: extracting and characterizing significant places from mobility data. In: 6th IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2011, Providence, RI, USA, 23–28 October 2011, pp. 161–170. IEEE Computer Society (2011). https://doi.org/10.1109/VAST.2011.6102454
3. Andrienko, N.V., Andrienko, G.L., Barrett, L., Dostie, M., Henzi, S.P.: Space transformation for understanding group movement. IEEE Trans. Vis. Comput. Graph. **19**(12), 2169–2178 (2013). https://doi.org/10.1109/TVCG.2013.193

4. Ankerst, M., Breunig, M.M., Kriegel, H., Sander, J.: OPTICS: ordering points to identify the clustering structure. In: Delis, A., Faloutsos, C., Ghandeharizadeh, S. (eds.) SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, 1–3 June 1999, Philadelphia, Pennsylvania, USA, pp. 49–60. ACM Press (1999). https://doi.org/10.1145/304182.304187
5. Barth, D., Bellahsene, S., Kloul, L.: Mobility prediction using mobile user profiles. In: MASCOTS 2011, 19th Annual IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, Singapore, 25–27 July 2011, pp. 286–294. IEEE Computer Society (2011). https://doi.org/10.1109/MASCOTS.2011.57
6. Ester, M., Kriegel, H., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E., Han, J., Fayyad, U.M. (eds.) Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA, pp. 226–231. AAAI Press (1996). http://www.aaai.org/Library/KDD/1996/kdd96-037.php
7. Guidotti, R., Nanni, M.: Crash prediction and risk assessment with individual mobility networks. In: 2020 21st IEEE International conference on mobile data management (MDM), pp. 89–98. IEEE (2020)
8. Jeung, H., Yiu, M.L., Zhou, X., Jensen, C.S.: Path prediction and predictive range querying in road network databases. VLDB J. **19**(4), 585–602 (2010). https://doi.org/10.1007/S00778-010-0181-Y
9. Kruskal, J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Am. Math. Soc. **7**(1), 48–50 (1956)
10. Landi, C., Guidotti, R., Nanni, M., Monreale, A.: The trajectory interval forest classifier for trajectory classification. In: SIGSPATIAL/GIS, pp. 67:1–67:4. ACM (2023)
11. Landi, C., Spinnato, F., Guidotti, R., Monreale, A., Nanni, M.: Geolet: an interpretable model for trajectory classification. In: IDA. Lecture Notes in Computer Science, vol. 13876, pp. 236–248. Springer (2023)
12. Larcom, S., Rauch, F., Willems, T.: The benefits of forced experimentation: striking evidence from the London underground network. Q. J. Econ. **132**(4), 2019–2055 (2017)
13. de Montjoye, Y.A., Shmueli, E., Wang, S.S., Pentland, A.S.: openPDS: protecting the privacy of metadata through safeanswers. PLOS One **9**(7), 1–9 (2014). https://doi.org/10.1371/journal.pone.0098790
14. Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabási, A.L.: Returners and explorers dichotomy in human mobility. Nat. Commun. **6**(1), 8166 (2015)
15. Ramer, U.: An iterative procedure for the polygonal approximation of plane curves. Comput. Graph. Image Process. **1**(3), 244–256 (1972). https://doi.org/10.1016/S0146-664X(72)80017-0
16. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, Colorado, USA, 3–6 September 1996, pp. 336–343. IEEE Computer Society (1996). https://doi.org/10.1109/VL.1996.545307
17. Trasarti, R., Pinelli, F., Nanni, M., Giannotti, F.: Mining mobility user profiles for car pooling. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1190–1198 (2011)
18. Wang, R., Zhang, X., Li, N.: Zooming into mobility to understand cities: a review of mobility-driven urban studies. Cities **130**, 103939 (2022)

19. Zhao, X., Pi, D., Chen, J.: Novel trajectory privacy-preserving method based on prefix tree using differential privacy. Knowl. Based Syst. **198**, 105940 (2020). https://doi.org/10.1016/J.KNOSYS.2020.105940
20. Zhou, H., Hirasawa, K.: Spatiotemporal traffic network analysis: technology and applications. Knowl. Inf. Syst. **60**(1), 25–61 (2019). https://doi.org/10.1007/S10115-018-1225-7