

---

# METHODOLOGY FOR MINING, DISCOVERING AND ANALYZING SEMANTIC HUMAN MOBILITY BEHAVIORS

---

**Clement Moreau, Thomas Devogeole, Veronika Peralta, Cyril de Runz**  
 University of Tours, France  
 firstname.lastname@univ-tours.fr

**Laurent Etienne**  
 LABISEN, Yncrea Ouest, France  
 laurent.etienne@yncrea.fr

December 22, 2020

## ABSTRACT

Various institutes produce large semantic datasets containing information regarding daily activities and human mobility. The analysis and understanding of such data are crucial for urban planning, socio-psychology, political sciences, and epidemiology. However, none of the typical data mining processes have been customized for the thorough analysis of semantic mobility sequences to translate data into understandable behaviors. Based on an extended literature review, we propose a novel methodological pipeline called SIMBA (Semantic Indicators for Mobility and Behavior Analysis), for mining and analyzing semantic mobility sequences to identify coherent information and human behaviors. A framework for semantic sequence mobility analysis and clustering explicability based on integrating different complementary statistical indicators and visual tools is implemented. To validate this methodology, we used a large set of real daily mobility sequences obtained from a household travel survey. Complementary knowledge is automatically discovered in the proposed method.

## 1 Introduction

It is becoming increasingly important to have a good understanding of human mobility patterns in many fields, such as transportation [10], social and political sciences [46, 13], epidemiology [60, 16], and smart city design [55]. For the latter, the ability to model urban daily activities correctly for traffic control, energy consumption, and urban planning [9] has a critical impact on human quality of life and the everyday functioning of cities. To inform policy makers regarding important projects, such as planning new metro lines, managing traffic demand during large events, or constructing shopping malls, we require reliable models of urban travel demand. Such models can be constructed from censuses, household travel surveys, or simulations that attempt to learn about human behaviors in cities using data collected from location-aware technologies [37, 56]. The development of generative algorithms that can reproduce and aid in understanding human mobility behaviors accurately is fundamental for designing smarter and more sustainable infrastructures, economies, services, and cities [11]. Typically, mobility analysis focuses on spatiotemporal analysis and the properties of human movement [8]. Pioneering works have highlighted the characteristics, regularities, and predictability of human mobility [7, 12, 29, 67, 68, 36].

Recently, a major challenge in machine learning and clustering methods has been the ability to explain models both for both practical and ethical purposes [31]. Explanation technologies and techniques are immensely helpful for companies that wish to improve the management and understanding of customer needs. They are also important for improving the openness of scientific discovery and progress of research. The need for clear and interpretable results and models is increasing, particularly for black-box algorithms, where data are huge and complex, and for methods with many parameters. Interpretability is crucial for testing, observing, and understanding the differences between models. Therefore, data comprehension also enhances the learning and/or exploration process in terms of validity.

In this paper, we propose transposing studies on human movement analysis into the semantic domain to learn and understand human activities. We focus on the analysis of semantic sequences of daily activities and attempt to learn and understand the properties of semantic mobility to extract consistent behaviors from a real human mobility dataset. In summary, this paper provides the following main contributions:

- A methodological pipeline called *Semantic Indicators for Mobility and Behavior Analysis* (SIMBA) is proposed to extract, analyze, and summarize coherent behaviors in semantic sequences of mobility data.
- We propose a framework for semantic sequence mobility analysis and clustering explicability integrating state-of-the-art indicators.
- A case study, from a real-world dataset to the extraction of understandable behaviors, illustrating the applicability of our proposal.

To the best of our knowledge, such a methodology for mining and interpreting clusters of human semantic mobility behaviors has not been proposed previously. Additionally, our methodology is generic and can be applied to any type of data representing a sequence of semantic symbols (e.g., activities, points of interest (POIs), web pages, and music in playlists). The remainder of this paper is organized as follows. Section 2 presents related work on human mobility indicators and methods for behavior extraction. In Section 3, we introduce some preliminary definitions and present an overview of our approach. We then discuss the design, statistics, and analytical methods used for behavior extraction from semantic sequences. Section 4 is dedicated to data description and global analysis of the target dataset. Section 5 discusses the extraction and characterization of behaviors using a clustering method and the explicability of discovered patterns. We also discuss results in this section. Finally, Section 6 concludes this article.

## 2 Related work

Questions regarding the extraction of mobility behaviors and comprehension of discovered patterns lie on the intersection of three main subjects, namely the study of human mobility properties, methods for comparing two semantic mobility sequences (i.e., two sequences of daily activities), and the explicability and interpretability of abstruse machine learning models. Therefore, in this section, we summarize major studies on human mobility characteristics as a basis for the requirements of similarity measures between mobility sequences. An extensive review of similarity measures and their properties is presented in the second subsection. The third subsection discusses clustering methods based on arbitrary distance matrices for automatically extracting groups of similar individuals. Finally, we discuss tools for human mobility analysis from the literature and commonly used indicators for describing semantic mobility sequences, as well as the explainability of black-box models, to understand and infer behaviors.

### 2.1 Human mobility properties

Numerous studies on human mobility have shown remarkable heterogeneity in travel patterns that coexist with a high degree of predictability [3]. In other words, individuals exhibit a large spectrum of mobility ranges while repeating daily schedules that are dictated by routine. González et al. analyzed a nation-wide mobile phone dataset and found that human trajectories exhibit a high degree of temporal and spatial regularity. Each individual is characterized by a time-independent characteristic travel distance and has a significant probability of returning to a few frequently visited locations [29]. In particular, the authors highlighted the following points. (i) According to Brockmann et al. [12], the travel distances of individuals follow a power-law behavior distribution. (ii) The radius of gyration of individuals, which represents their characteristic travel distance, follows a truncated power law. Song et al. [67] observed mobile phone data and determined that the waiting times of individuals (i.e., times between two moves) are characterized by a power-law distribution, confirming the results presented by Barabási [7]. Additionally, in [68], Song et al. analyzed the movements of individuals based on the Lempel-Ziv algorithm [40] and calculated a value of 93% potential predictability for user mobility, which demonstrates that a significant portion of predictability is encoded in the temporal orders of visitation patterns. Additionally, despite significant differences in travel patterns, the variability in predictability is weak and largely independent of the distances users cover on a regular basis. This study was continued by Texeira et al. [71], who demonstrated that the entropy of a mobility sequence can be estimated using two simple indicators, namely regularity and stationarity, indicating that trivial indicators can capture the complexity of human mobility.

When considering patterns in human mobility, particularly movements within a single day or week, it is essential to distinguish locations based on their importance. As mentioned previously, people exhibit periods of high-frequency trips followed by periods of lower activity and a tendency to return home on a daily basis. Therefore, most daily and weekly trajectories will start and end at the same location [8]. One method for quantifying the importance of locations is to rank locations. In [67], location ranking was performed for mobile users based on the numbers of times their positions were recorded in the vicinities of the cell towers covering their locations. It was found that visitation frequency follows a Zipf law. Another method of distinguishing locations is to construct individual mobility patterns in the form of a network. Schneider et al. [66] used data from both mobile phone users and travel survey respondents to construct weekday mobility networks for individuals. These profiles consisted of nodes for updating visited locations and edges for modeling trips between locations. Daily networks were only constructed for weekdays to identify topological

patterns in mobility during a typical day. It was determined that approximately 90% of the recorded trips made by all users could be described using only 17 daily networks. Another important point is that all of the identified networks contained strictly less than seven nodes and most of the networks exhibited oscillations, which are represented by cyclic links between two or more nodes. This result suggests that these motifs represent the underlying regularities in our daily movements and are useful for the accurate modeling and simulation of human mobility patterns.

## 2.2 Approaches to semantic mobility sequence mining

In mobility mining, two main approaches coexist with distinct goals: sequence pattern mining and clustering methods. The former extracts subsequences of frequent items from trajectories [28, 77, 74, 23] to represent an aggregated abstraction of many individual trajectories sharing the property of visiting the same sequence of places with similar travel and visit times. The latter constructs clusters of similar sequences by comparing pairs using a similarity measure. Each cluster represents a coherent behavior and shares mobility features according to similarity measure properties.

Although sequence pattern mining methods are efficient for mining regular fragments and are easy to interpret, they are unsuitable for assessing similarities between individuals, meaning they cannot be used to extract representative groups according to their activities accurately. In general, clustering methods are superior for comparison, classification, and grouping tasks. Therefore, in the remainder of this section, we review related work on clustering processes for mining mobility behavior. Specifically, we focus on difficulties and solutions associated with the choice of a similarity metric between semantic mobility sequences.

### 2.2.1 Similarity measures

Many similarity measures have been proposed or adapted for comparing sequences of symbols, specifically spatial trajectories (e.g., Euclidean, LCSS [72], DTW [39], EDR [15] and Fréchet [4]). Most of these measures have been adapted for semantic human mobility to compare sequences of routine activities or location histories [45, 36, 47]. Table 1 summarizes the measures reviewed, which can be classified into two broad categories: measures based on counts of different attributes between sequences (Att) and measures based on edit distances, which measure the cost of the operations required to transform one sequence into the other (Edit).

Because trajectories are complex objects based on their multidimensional aspects, the construction and analysis of similarity measures remains a challenging task, underlying by [22], and few works have successfully handled multiple dimensions. In [25] and [42], two similarity measures for multidimensional sequences called MSM and SMSM, respectively, were defined based on the aggregation of matching functions controlled by weighting distances defined for each dimension of a sequence. These multidimensional similarity measures can embed the richness inherent to mobility data, but require many parameters and thresholds for initialization. This complexity makes it difficult to visualize and interpret the resulting similarity scores.

Most previous proposals focus on unidimensional semantic sequences. One method for comparing semantic sequences is to represent them as vectors. Such a representation is particularly interesting because it allows the use of a whole family of distance measures that are well-defined metrics, such as the inner product and Euclidean distance. In [18], Elzinga and Studer represented sequences as vectors in the inner product space and proposed a context metric called SVRspell that focuses on duration and similarity. However, their representation has an exponential space complexity of  $\mathcal{O}(|\Sigma|^n)$  where  $|\Sigma|$  is the size of the alphabet of symbols and  $n$  is the size of the sequence.

Jiang et al. also represented daily activity sequences as vectors. They defined the space of an individual's daily activity sequence by dividing the 24 h in a day into five minutes intervals and then used the activity in the first minute of every time interval to represent an individual's activity during that five minutes timeframe. Principal component analysis was then used to extract appropriate Eigen activities and calculate the Euclidean distances between them [36]. It should be noted that, in this previous work, time slots are the kernel level for comparison in the sense of two individuals with same activities but practised at different times will be evaluated as strongly dissimilar. We call this type of approach a *time-structural approach*. It is effective to group individuals based on they allocate time to different activities, but a major problem with this approach is that two trajectories composed of the same activities practiced at different times will have no similarity, resulting in extreme sensitivity to time.

To overcome this time issue, other studies have reused measures from optimal matching (OM) methods [70] such as the edit distance family (e.g., Levenshtein), LCSS, and DTW. These methods measure the dissimilarity between two sequences  $S_1$  and  $S_2$  as the minimum total cost of transforming one sequence (e.g.,  $S_1$ ) into the other (e.g.,  $S_2$ ) using indels (insertions, deletions) or substitutions of symbols. Each of these operations has a cost that can vary with the states involved. In this manner, depending on the choice of costs applied, groups of individuals can be created differently.

Table 1: Description of main similarity measures for semantic sequences

Measure	Type		Description
	Att.	Edit.	
MSM [25], SMSM [42]	×		Aggregation of matching functions of each dimension.
SVRspell [18]	×		Based on number of matching subsequences and weighted by the length of subsequences involved.
Jiang et al. [36]	×		Euclidean distance between appropriate eigen activities.
Hamming	×		Sum of mismatches with similarity between elements.
DHD [43]	×		Sum of mismatches with positionwise state-dependent weights.
Levenshtein distance [44, 73]		×	Minimum sum of edit costs to turn $S_1$ into $S_2$ .
CED [52]		×	OM with costs weighted by edit position and symbols nearby.
Trate (TDA) [63]		×	OM with costs based on transition rates.
OMSlen [34]		×	OM with costs weighted by symbol length.

Table 2: Properties of main similarity measures for semantic sequences

Measure	Metric	Properties					
		T. warp	Ctxt	Permut.	Rep.	Sim.	Multi. dim
MSM [25], SMSM [42]		×					×
SVRspell [18]	×	×	×		×	×	
Jiang et al. [36]	×						
Hamming	×†					×‡	
DHD [43]			×				
Levenshtein distance [44, 73]	×†	×				×‡	
CED [52]	×†	×	×	×	×	×	
Trate (TDA) [63]		×	×			×	
OMSlen [34]	×	×			×		

† Depends if costs fulfil the triangle inequality and/or parameters.

‡ By default discrete metric  $\rho(x, y) = \begin{cases} 0 & x = y \\ 1 & \text{else} \end{cases}$

## 2.2.2 OM methods, setting cost and mobility behavior

A major challenge in OM-based methods is setting operation costs. This is a particularly difficult problem in social science [1, 35]. There are essentially three main strategies for choosing operation costs: (i) theory-based cost [70], which determines costs based on theoretical grounds and a priori knowledge; (ii) feature-based cost, which specifies a list of state attributes on which we wish to evaluate the closeness between states using a similarity measure such as the Gower index [30] or Euclidean distance; and (iii) data-driven cost [63], which assigns a cost that is inversely proportional to the transition rates observed in the dataset. A well-known example of the latter strategy is Dynamic Hamming Distance (DHD) [43] where the substitution costs at position  $t$  are obtained by the transition rates cross-sectionally observed between  $t - 1$  and  $t$  and between  $t$  and  $t + 1$ . This method is very effective at generating abnormal sequences and outliers. However, based on its construction, DHD has strong time sensitivity and the number of transition rates that must be estimated is very high, potentially leading to overfitting. Finally, there is an additional type of strategy called (iv) ontology-based cost (utilized in [52] that is derived from (i) and (ii). This approach infers costs based on taxonomies (or ontologies) and similarity measures in knowledge graphs [78].

An additional difficulty in setting operation costs lies in the context of sequences or, in other words, considering the symbols in the sequences. As pointed out in [29, 68, 3], human mobility has a high degree of regularity. Several approaches have been developed to take advantage of this regularity. In [34], the OMSlen method was proposed to reduce the cost of operations for repeating symbols, which is particularly useful for mobility sequence mining. Moreau et al. [52] proposed reducing the costs of edit operations for symbols that are similar and/or already present in a sequence and close to the edited position. One consequence of this method is that repetitions of nearby similar symbols and permutations have lower costs, making it a *compositional comparison approach*. This method can bring together sequences with similar contents by allowing for some temporal distortions and repetitions.

Table 3: Complexity and parameters of main similarity measures for semantic sequences

Measure	Complexity	Parameters		
		Subs	Indels	Other
MSM [25], SMSM [42]	$\mathcal{O}(n \times p)$			Set of distances $\mathcal{D}$ ; weight vector $w$ ; threshold vector $maxDist$
SVRspell [18]	$\mathcal{O}( \Sigma ^{\max(n,p)})$			Subsequence length weight $a$ ; symbol duration weight $b$
Jiang et al. [36]	$\mathcal{O}(n \times p)$			Number of activities
Hamming	$\mathcal{O}(n)$	Single, User <sup>‡</sup>		
DHD [43]	$\mathcal{O}(n)$	Data		
Levenshtein distance [44, 73]	$\mathcal{O}(n \times p)$	Single, User <sup>‡</sup>	Single	
CED [52]	$\mathcal{O}(n \times p \times \max(n, p))$	Ontology	Auto	Ontology; Context function $f_k$ ; Context weight $\alpha$
Trate (TDA) [63]	$\mathcal{O}(n \times p)$	Data	Single	Transition lag $q$
OMSlen [34]	$\mathcal{O}(n \times p)$	User	Multiple	Symbol length weight $h$

<sup>‡</sup> If user specifies a similarity measure.

Based on the information in [70], a summary of measure properties is presented in Table 2. The column “Metric” denotes measures based on mathematical calculations of distances. “T. warp” denotes measures allowing time warping when comparing sequences. “Ctx” denotes measures that consider the context of a sequence to define cost. “Permutat.” indicates that permutations are allowed with a lower cost and “Rep.” indicates that repetitions are cheaper. Finally, “Sim” denotes measures that consider a similarity function between symbols and “Multi. dim” denotes measures that handle multidimensional sequences.

Table 3 presents the computational complexity and some details regarding the parameters of each method. In the “Complexity” column,  $n$  and  $p$  denote the lengths of compared sequences. It should be noted that for Hamming-family measures, the sequences must have the same length ( $n$ ). The column “Parameters” contains the necessary tuning parameters and cost strategies for OM measures. In the “Subst” columns, an entry of “User” indicates that the costs are set by the user through a theory- or feature-based strategy. An entry of “Data” denotes a data-driven method and “Ontology” refers to an ontology-based strategy. Finally, the “Indels” column indicates whether there is a single state-independent indel cost, denoted as “Single” state-dependent user-defined indel costs, denoted as “Multiple” or indel costs that are automatically set by the measure itself, denoted “Auto”.

### 2.2.3 Clustering methods

The extraction of behaviors from a dataset is a process that is typically performed using unsupervised machine learning methods. Clustering methods are based on similarity measures such as those described in the previous subsections and are widely used for the discovery of human behaviors, particularly in sequences of mobility data [36, 75, 56].

However, the topologies created by similarity measures for semantic sequences are difficult to apply. In particular, for OM methods, the axioms of the metric spaces are, usually, not hold. A pairwise comparison of semantic sequences results in a distance matrix that is used as an input for a clustering process. To the best of our knowledge, the clustering algorithms that are able to handle arbitrary distances (not necessarily metrics) are PAM (or k-medoid) [59], hierarchical clustering [38], density clustering (DBSCAN [19], OPTICS [5]) and spectral clustering [54], each of which proposes different hypotheses regarding cluster topology.

According to the similarity measure and representation of sequences, dimensionality reduction methods can be applied to extract primary dimensions [36]. However, commonly used methods such as PCA can only be used for Euclidean spaces in practice. Alternatively, methods such as UMAP [49], facilitate the reduction of a complex topology defined by an arbitrary metric into a low Euclidean space, which facilitates the visualization of clustering results and the use of other clustering methods, such as those requiring a Euclidean space, including k-means. Additionally, UMAP offers superior preservation of the data global structure, fewer hyperparameters for tuning, and better speeds than previous techniques such as t-SNE [48].

Therefore, the advantages of these clustering techniques is that they can be used with arbitrary distances, meaning they can be paired with any of the measures discussed in Section 2.2.1 to implement a clustering module.

Table 4: Indicators for explainability and analysis of semantic mobility sequences and behaviors in a dataset

Techniques	Refs	Description
<b>Frequency distribution</b>		
Length distribution		Frequency distribution of sequences length in the dataset.
State distribution		Frequency distribution for each symbol $x$ in the sequences in whole dataset.
<b>Transition</b>		
Origin-Destination matrix		Number of transitions from a state (i.e. symbol) $x_i$ to $x_j$ .
Daily pattern	[66]	Network representation of sequence. Each edge $(x_i, x_j)$ represent a transition from $x_i$ to $x_j$ .
<b>Disorder</b>		
Entropy	[67, 40]	Level of “information”, “surprise”, or “uncertainty” inherent of a variable’s possible outcomes. For sequences, the entropy also consider temporal patterns.
Predictability	[67]	Probability that an appropriate predictive algorithm can predict correctly the user’s future whereabouts.
Distinct symbols	[71]	Number of distinct symbols in the sequence.
<b>Statiscal dependance measures</b>		
Association rules	[2]	Relation, based on measures of interestingness, between two or more variables in a dataset.
Pearson residuals	[32]	Measure of the departure of the independence between two variables.
<b>Centrality</b>		
Mode		Element with the highest frequency in the dataset.
Medoid		Element which minimizes the distance to all elements in the dataset.
<b>Scattering and outliers</b>		
Diameter and Radius		Geometrical interpretation of the distances between elements in the dataset.
Distance distribution		Distribution of the distance between in the dataset or subset of it (e.g. cluster).
Silhouette	[65]	Quality score of clustering. Measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation)
UMAP	[49]	Dimensional reduction. Visualization of complex elements in 2D Euclidean spaces with a preservation of local topology.

### 2.3 Analysis tools supporting mobility mining

Data mining and statistical learning techniques are powerful analysis tools for understanding urban mobility [55]. Several works have proposed frameworks and tools for supporting sequence analysis and mobility mining. The most relevant tools are briefly described in this subsection.

One of the first and well-known frameworks for mobility knowledge discovery is M-Atlas [27], which provides complete functionalities for mobility querying and data mining centered around the concept of trajectories.

Recently, [57] proposed a statistical python library for mobility analysis called *Scikit-Mobility*. Scikit-Mobility enables to load, clean, process and represent mobility data and analyze these by using the common mobility measures. However, to the best of our knowledge, there is no framework oriented semantic mobility mining. Some toolboxes provide partial statistical support for mobility analysis (mainly oriented to spatial mining). One can see [61] for a review of R libraries and [57] for Python.

Based on TraMineR [26] functionalities, the geovisualization environment eSTIME [51] allows the representation of semantic daily mobility information with spatio-temporal content.

Despite the availability of such decision support tools and reporting systems, abstracting and analyzing the main characteristics of a group of semantic sequences and explaining why they are clustered together remains a challenge open problem. In particular, while the interpretability and explainability of mining methods are hot research topics, most methods are limited to a specific problem or domain. To the best of our knowledge, there have only been a few studies on providing a methodology for understand mobility mechanisms in clusters of semantic sequences. The most relevant work is that described in [36], where a K-means clustering method based on a time-structured similarity measure was applied to daily mobility sequences. The authors defined eight clusters corresponding to predefined social demographic variables. Cluster analysis was mainly performed based on sequence index plots, state distributions, and the proportion of social demographic characteristics in each cluster. The Silhouette index [65] was used to control clustering validity. Although this analysis provides a starting point for understanding the typical behaviors within a cluster, it is incomplete

and fails to qualify how consistent the elements in a cluster are with the cluster description (e.g., most extreme elements in a cluster and entropy of sequences in a cluster), as well as the topologies formed by mobility sequences (e.g., daily patterns). These aspects of explainability must be retained and enriched to provide a set of indicators allowing us to understand the globality and diversity of all the elements in a cluster, as well as what makes a cluster coherent.

Techniques that attempt to explain complex machine learning methods are becoming increasingly popular. For example, the LIME technique [62] attempts to explain the predictions and results of black-box machine learning techniques in an interpretable and faithful manner by training an interpretable model based on local results. Similarly, Guidotti and al. [31] proposed some techniques and methods such as association and decision rules, and prototype selection elements (e.g., medoids and diameters) to explain black-box systems to make their results more interpretable. In line with these techniques, we believe that the elaboration of indicators is a crucial point for understanding machine learning models.

Table 4 presents a summary of the different indicators used in state-of-the-art methods that can be used to explain semantic mobility sequences. The indicators are structured into categories corresponding to different perspectives of exploring and explaining data. We let  $X = \langle x_1, x_2, \dots, x_n \rangle$  denote a sequence of symbols constructed from an alphabet  $\Sigma$  and let  $f$  denote the frequency function.

In this paper, we address the problem of knowledge extraction from human activity sequences to develop models of mobility behaviors. To this end, we reuse many of the techniques introduced in this section and propose several new methods to mine and qualify semantic sequences.

### 3 Methodology

This section details the proposed methodology for the analysis of semantic mobility sequences. First, we summarize the methodological pipeline presented this paper, including the nature of the dataset, selected statistical analysis methods, indicators, and clustering process. The second subsection is dedicated to the enrichment and representation of semantic mobility sequences and the third subsection introduces the indicators used for semantic mobility sequence analysis. The fourth subsection discusses the clustering process and corresponding similarity measure, namely Contextual Edit Distance (CED), as well as a hierarchical clustering process. Finally, the fifth subsection describes the methodology used for cluster analysis and the extraction of semantic mobility behaviors.

#### 3.1 The SIMBA methodological pipeline

Semantic mobility sequences are complex data based on their nature and properties, as discussed in Section 2. However, daily mobility has a high degree of regularity with many repetitions of activities. Based on these characteristics, Figure 1 presents the SIMBA methodology based on the strengths of previous method (also discussed in Section 2). It consists of five steps labeled as (a), (b), (c), (d), and (e).

In the first step, semantic data are enriched using an ontology to facilitate the comparison of concepts based on any similarity measure that can be adapted to knowledge graphs and data with different levels of granularity.

In the second step, we compute some global statistics to understand and analyze the data. These statistics are selected from the indicators introduced in Table 4 and provide a complementary analysis of sequences in terms of their contents (frequency distribution), networks (transition), central behaviors (centrality), and degrees of disorder (entropy). Based on this complementarity, we can explain data from different perspectives. Additionally, the statistics highlight the different patterns that mobility sequences follow (visitation frequency, daily patterns, origin-destination matrices, sequence lengths, predictability) [67, 66, 67]), provide a preliminary overview of the data, and facilitate quality control.

The third step focuses on the clustering of semantic sequences, which groups sequences representing similar moving behaviors. The main challenge in this step is the comparison of semantic sequences, specifically the selection of a similarity measure to support such comparisons and adapt to specific business needs. In this study, we used the CED similarity measure [52]. This measure extends edit distance by adapting a cost computation for typical mobility characteristics, such as redundancies, repetitions [68] and cycles [66]. A pairwise comparison of semantic sequences yields a distance matrix, which is then used in the clustering process. Section 2.2.3 summarizes various approaches to sequence clustering. These clustering algorithms are based on different assumptions regarding cluster topology and can all be used in this step. However, because the topology of the semantic sequence space is difficult to comprehend, in this study, we visualized it using a dendrogram generated from a hierarchical clustering process.

The output of this step is a set of clusters of semantic sequences that represent similar behaviors. SIMBA is a modular methodology in which the similarity measure and clustering algorithm proposed in step (c) can be replaced with any of the other techniques discussed in Section 2.

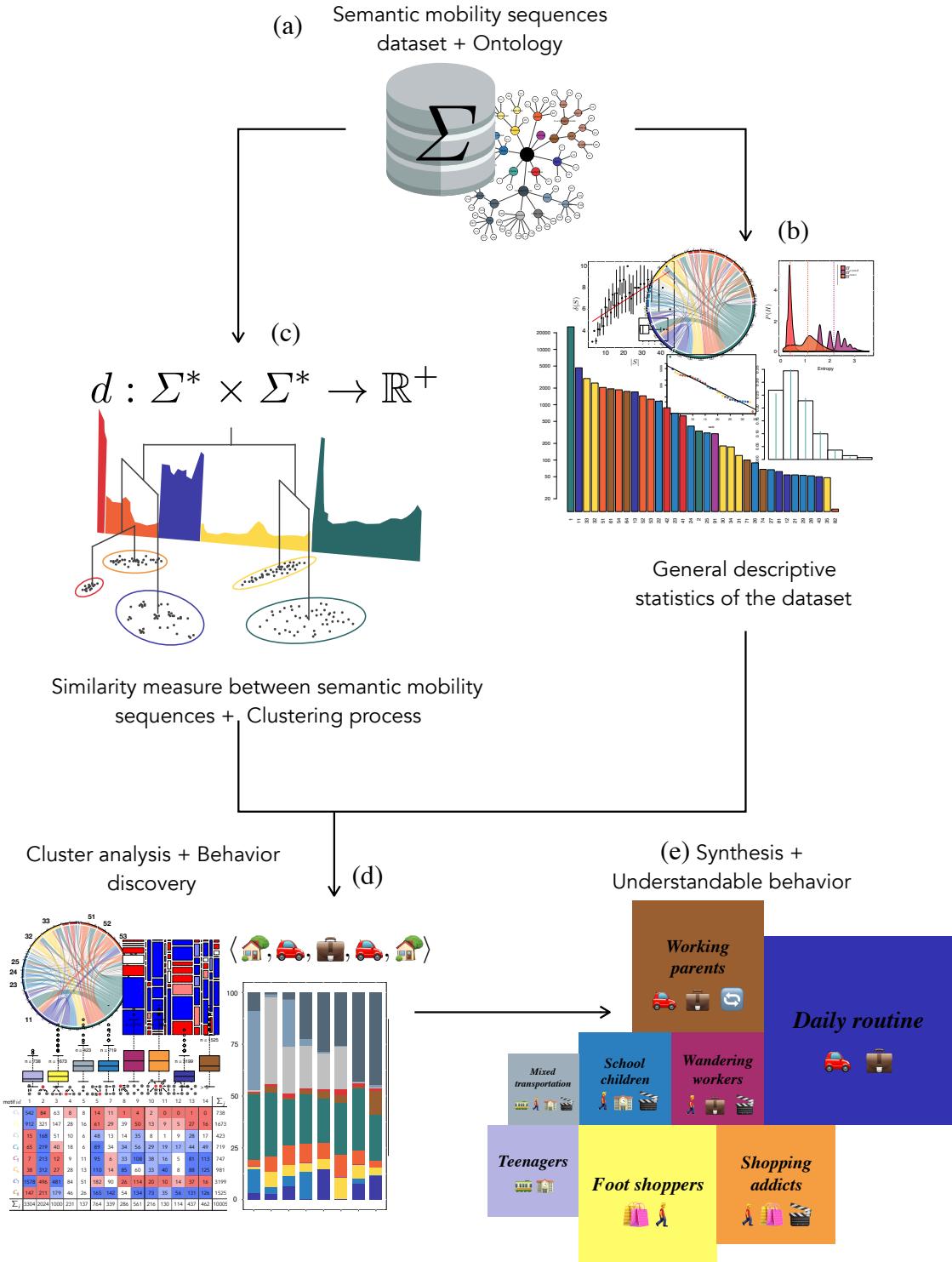


Figure 1: Overview of the simba methodology : (a) Given semantic mobility sequences dataset and an ontology of activities (b) General descriptive statistics and indicators (c) Clustering process (d) Analysis of semantic mobility clusters for behavior discovering (e) Synthesis of each behavior in an understandable way

Step (d) computes additional statistics for each cluster to extract and understand the specific characteristics that constitute mobility behaviors. The statistical and data visualization indicators partially replicate those used in the overall analysis in step (b), but are enhanced with significance tests to determine which are typical characteristics in terms of activities, patterns, and sizes in each cluster. Additionally, these indicators are studied in combination with clustering centrality indicators (centroid, mode, diameter, cluster variance) and quality measures (i.e., Silhouette score [65]), which measures intra-cluster and inter-cluster distances). This step can also be used to identify outliers.

Finally, step (e) summarizes the main characteristics of clusters to label them in terms of mobility behaviors. A graphical summary concludes the pipeline and yields an easy and understandable way to discover mobility behaviors.

The remainder of this section precisely describes each step of the SIMBA methodology.

### 3.2 Enrichment and representation of semantic mobility sequences

Let  $\Sigma$  be a set of concepts that represent daily activities (see Section 4 and Table 6). We define semantic mobility sequence as follows.

**Definition 1 (Semantic sequence)** *Given a human  $h$ , their semantic sequence<sup>1</sup>  $S$  is an ordered sequence of activities  $\langle x_1, x_2, \dots, x_n \rangle$  such as  $\forall k \in \llbracket 1, n \rrbracket, x_k \in \Sigma$  and for  $i < j, x_i$  predates  $x_j$ . Additionally, we consider that symbols are not repeated consecutively (i.e.,  $\forall k \in \llbracket 1, n-1 \rrbracket, x_k \neq x_{k+1}$ ).*

*Intuitively, such a sequence indicates that  $h$  performed activity  $x_1$ , then  $x_2$ , and finally  $x_n$ .*

To compare the symbols in  $\Sigma$ , we must introduce a partial order to the set. For this purpose, we construct a knowledge graph between all concepts in  $\Sigma$ .

**Definition 2 (Knowledge graph)** *Let  $\Sigma$  be a set of concepts such that  $\exists \text{root} \in \Sigma$ . A knowledge graph is a connected and directed acyclic graph  $G = (\Sigma, E)$  with  $E \subset \Sigma \times \Sigma$  where  $(x, y) \in E$  iff the concept  $x$  (meronym) contains semantically the concept  $y$  (holonym) and  $\forall (x, y) \in E, y \neq \text{root}$ . Such a knowledge graph is called a meronymy.*

*In the resulting graph of concepts, for any two concepts  $x, y \in \Sigma$ , we let  $LCA(x, y)$  denote the last common ancestor of  $x$  and  $y$ .  $d(x)$  denotes for  $x$ 's depth (i.e., its minimal distance from the root node).*

Additionally, knowledge representations such as an *is-a* taxonomy can induce a partial order on  $\Sigma$ . This classification of  $\Sigma$  allows us to define similarity measures on its elements. Many similarity measures have been proposed for knowledge graphs (see [78] for a survey). In the remainder of this paper, we use the Wu-Palmer similarity measure [76], which is defined a  $sim_{WP} : \Sigma \times \Sigma \rightarrow [0, 1]$ . This is a well-established state-of-the-art measure that accounts for both the depth of the concepts in an ontology and their closest ancestors, and is normalized:

$$sim_{WP}(x, y) = \frac{2 \times d(LCA(x, y))}{d(x) + d(y)} \quad (1)$$

Moreover, thanks to the hierarchical representation of activities, data can be analyzed at different aggregation levels, similar to online analytical processing analysis. For example, the activities of shopping in a mall and shopping in a marketplace can be aggregated into a single higher-level shopping activity. Intermediate nodes in a meronymy are useful for such aggregations.

### 3.3 Semantic mobility sequence dataset analysis

Semantic sequence data are difficult to analyze based on their combination of temporal dimensions (i.e., temporal order of activities) and semantic dimensions. As discussed in Section 2.1, human mobility semantic sequences tend to follow statistical laws. Frequent visitation items induce repetitions of activities that can be modeled using a Zipf law. Sequences are mainly structured by a few networks of daily patterns and are characterized by low entropy and a Poisson distribution for their length.

Therefore, to ensure the quality of a dataset in terms of the aforementioned properties and obtain a preliminary understanding the data, based on Table 4, we propose complementary statistical indicators that facilitate the global analysis of a set of semantic sequences. The selected indicators are listed in Table 5. Although this study focused on mobility sequences, the proposed methodology is generic and can be used for analyzing any type of semantic sequence dataset.

---

<sup>1</sup>Considering our use case, we use indistinctly semantic sequence, semantic mobility sequence or mobility sequence terms.

Table 5: Retained indicators for semantic mobility sequences analysis

Id	Techniques	Visualization methods	Used for		Example
			All dataset	Clusters	
<b><i>Frequency distribution</i></b>					
1	Length distribution	Histogram, box plot	×	×	Figs. 4, 11
2	State distribution	Histogram, stack plot	×	×	Figs. 3, 12
<b><i>Transitions</i></b>					
3	Origin-Destination matrix	Chord diagram	×	×	Fig. 5
4	Daily patterns	Network and histogram	×	×	Fig. 6
<b><i>Disorder</i></b>					
5	Entropy	Density plot	×		Fig. 8
6	Predictability	Density plot	×		Fig. 8
7	Distinct symbols	Box plot	×		Fig. 7
<b><i>Statiscal dependance measures</i></b>					
8	Pearson residuals	Mosaic diagram		×	Fig. 13
<b><i>Centrality</i></b>					
9	Mode	Emojis sequence		×	Tab. 8
10	Medoid	Emojis sequence		×	Tab. 8
<b><i>Scattering and outliers</i></b>					
11	Diameter and Radius	Table		×	Tab. 7
12	Silhouette	Table		×	Tab. 7

**Indicator 1 (Length distribution)** Frequency distribution of sequence length combined with a frequency histogram.

**Indicator 2 (State distribution)** Frequency distribution of activities  $x \in \Sigma$  inside the sequences of dataset combined with a frequency histogram. Using a log scale may be advisable in the field of human mobility.

Together, these two indicators provide a high-level overview of a sequence's content and length. However, they provide no information regarding the transitions or motifs in sequences. This analysis can be useful for estimating transition probabilities such as those used in DHD measures or for generating probabilistic models of flows. The resulting matrix can be visualized using a chord diagram. To this end, we incorporate the following additional indicators.

**Indicator 3 (Origin-destination Matrix)** Matrix  $T = \{t_{ij}\}$  in which each line/column represents an activity  $x \in \Sigma$ . The coefficient  $t_{ij}$  represents the number of transitions from activity  $x_i$  to activity  $x_j$ . Such a matrix can be visualized using a chord diagram.

**Indicator 4 (Daily pattern)** Frequency distribution of non-isomorph daily pattern graphs [66]. We compute this indicator using Algorithm 1:

---

**Algorithm 1:** Daily patterns frequency

---

**Data:** Dataset of semantic sequences  $\mathcal{D}$

**Result:** Dictionary  $\mathcal{G}$  of non-isomorph daily patterns graphs frequencies

$\mathcal{G} \leftarrow \emptyset$  ▷ Dictionary  $\mathcal{G}$  where keys are graphs and values are integers

▷ Construct the daily pattern graph of each sequence  $S \in \mathcal{D}$

**for**  $S \in \mathcal{D}$  **do**

$V_S \leftarrow \{x | x \in S\}$  ▷ Set of vertices

$E_S \leftarrow \{(x_i, x_{i+1}) | i \in [1, |S| - 1]\}$  ▷ Set of edges

$G_S \leftarrow (V_S, E_S)$

**if**  $\exists G \in \mathcal{G}.\text{keys}() | G \simeq G_S$  **then**

    ▷ If there is already exists a graph  $G$  isomorph to  $G_S$  in  $\mathcal{G}$

$\mathcal{G}[G] \leftarrow \mathcal{G}[G] + 1$  ▷ Increment the frequency of  $G$

**else**

$\mathcal{G}[G_S] \leftarrow 1$  ▷ Create it in  $\mathcal{G}$

**end**

**end**

---

It should be noted that the isomorphism test for the two graphs  $G$  and  $G'$  can be implemented using the Nauty algorithm [50].

Finally, to capture the degree of disorder in sequences and understand how studied sequences are both predictable and varied, we use the following three indicators developed in entropy studies.

**Indicator 5 (Entropy of a sequence)** The entropy of a sequence is defined in [68], where several types of entropy are given.

- The random entropy  $H^{rand} = \log_2 \delta(S)$ , where  $\delta(S)$  is the number of distinct activities in sequence  $S$ .
- The temporal-uncorrelated entropy  $H^{unc} = -\sum_{i=1}^{|S|} p(x_i) \log_2 p(x_i)$ , where  $p(x_i)$  is the historical probability that activity  $x_i$  was performed. This characterizes the heterogeneity of activities.
- The real sequence entropy  $H$  which depends on both, the frequency and the order of an activity in the sequence.

$$H(S) = - \sum_{S' \subset S} p(S') \log_2 p(S') \quad (2)$$

where  $p(S')$  is the probability of finding a particular ordered subsequence  $S'$  in the sequence  $S$ .

In practice,  $H$  is uncomputable for long sequences. Therefore, we use the following estimator  $H^{est}$  of  $H$  proposed in [40]:

$$H^{est}(S) = \left( \frac{1}{|S|} \sum_i \lambda_i \right)^{-1} \log_2 |S| \quad (3)$$

where  $\lambda_i = \text{argmin}_{k \geq 1} \{x_i \dots x_k \notin x_1 \dots x_{i-1}\}$  is the size of the smallest subsequence beginning at  $i$  and not contained in the and not contained in the range of 1 to  $i - 1$ .

Kontoyiannis et al. demonstrated that  $\lim_{|S| \rightarrow \infty} H^{est}(S) = H(S)$ , supplements can also be derived [68].

**Indicator 6 (Predictability)** The predictability  $\Pi$  that an appropriate algorithm can predict correctly the user's future whereabouts. Thanks to Fano's inequality, we can obtain an upper bound  $\Pi^{max}$  for  $\Pi$  [68].  $\Pi^{max}$  is obtained via the approximate resolution of the following equation:

$$H(S) = \mathcal{H}(\Pi^{max}) + (1 - \Pi^{max}) \log_2(|S| - 1) \quad (4)$$

where  $\mathcal{H}(x) = -x \log_2 x - (1 - x) \log_2(1 - x)$  is the binary entropy function.

**Indicator 7 (Distinct symbols)** *The frequency distribution of the number of distinct activities  $\delta$  in each sequence  $S$  in the dataset combined with a frequency histogram.  $\delta$  can also be studied in combination with the length  $|S|$  to uncover hidden regularities in a sequence.*

### 3.4 Clustering design for semantic sequences

To address the problem of clustering semantic mobility behaviors in a metropolitan area utilizing a compositional approach (i.e., “What does an individual do during a day?”), we use a combination of the CED measure and hierarchical clustering based on Ward’s criterion.

#### 3.4.1 CED

As discussed in Section 2.2.2, the distances in the edit distance family count the minimum costs of operations (e.g., modification, addition, deletion) required to transform one sequence into another. Such measures can be used to quantify the similarity between two semantic mobility sequences. However, as indicated in Section 2.1, human mobility sequences are characterized by redundancy of certain symbols, repetition [68], and cycles [66]. These features should be considered by adopting specialized distances.

Based on these observations, we proposed the use of the CED measure [53, 52], which is a generalization of edit distance for handling semantic mobility sequences. This measure incorporates the following factors:

1. *Context-dependent cost*: Edit cost depends on the similarity of nearby activities. The more similar and closer two activities are, the lower the cost of operations..
2. *Repetition*: Editing repeated nearby activities has a low cost.
3. *Permutation*: Similar and nearby activities can be exchanged with a low cost.

These three factors of CED are particularly suitable for mobility analysis. The fact that repetition and the editing of similar elements in a sequence has a low cost, similar to permutation, tends to group elements with activities with the same semantics while accounting for a flexible timeframe.

To achieve these advantages, the CED includes a modification of the cost operation function  $\gamma$  that generalizes the classical definition of edit distance and accounts for the local context of each activity in a mobility sequence.

Let a contextual edit operation be a quad tuple such that:

$$e = (o, S, x, k) \in \{\text{add}, \text{mod}, \text{del}\} \times \Sigma^n \times \Sigma \cup \{\varepsilon\} \times \mathbb{N}^*$$

where  $e$  is a transformation  $o$  of sequence  $S$  at index  $k$  using symbol  $x$ . Let  $E$  be the set of all possible contextual edit operations, the cost function  $\gamma : E \rightarrow [0, 1]$  for a contextual edit operations is defined as:

$$\gamma(e) = \alpha \times \ell(e) + (1 - \alpha) \left( 1 - \max_{i \in \llbracket 1, n \rrbracket} \{ \text{sim}(x, s_i) \times v_i(e) \} \right) \quad (5)$$

where:

- $\alpha \in [0, 1]$  is a contextual coefficient.  
If  $\alpha \rightarrow 0$ , then the cost will be strongly evaluated according to the near content at index  $k$  in the sequence being edited. If  $\alpha \rightarrow 1$ , then CED tends toward the Levenshtein Distance with substitution cost.
- $\ell(e) = \begin{cases} 1 - \text{sim}(s_k, x) & \text{if } o = \text{mod} \\ 1 & \text{else} \end{cases}$  is the cost function of Levenshtein Distance with substitution cost.
- $\text{sim} : \Sigma \times \Sigma \rightarrow [0, 1]$  is a similarity measure between two activities.
- $v(e) \in [0, 1]^n$  is a contextual vector that quantifies the notion of proximity between activities. Typically, the larger  $|i - k|$  is, the smaller  $v_i(e)$  is.

Let  $\mathcal{P}(S_1, S_2)$ , all the edit paths to transform a sequence  $S_1$  into  $S_2$ , the one-sided contextual edit distance from  $S_1$  to  $S_2$  noted  $\tilde{d}_{CED} : \Sigma^n \times \Sigma^p \rightarrow \mathbb{R}^+$  is defined such that:

$$\tilde{d}_{CED}(S_1, S_2) = \min_{P \in \mathcal{P}(S_1, S_2)} \left\{ \sum_{i=1}^{|P|} \gamma(e_i) \right\} \quad (6)$$

where  $P = (e_1, \dots, e_q) \in E^q$  is a vector of contextual edit operations.

The computation of Equation 6 is performed using dynamic programming and the Wagner-Fisher algorithm [73]. Finally,  $d_{CED} : \Sigma^n \times \Sigma^p \rightarrow \mathbb{R}^+$  is computed using the following equation:

$$d_{CED}(S_1, S_2) = \max \left\{ \tilde{d}_{CED}(S_1, S_2), \tilde{d}_{CED}(S_2, S_1) \right\} \quad (7)$$

### 3.4.2 Hierarchical clustering settings and validity

Hierarchical clustering algorithms have been widely applied to partition datasets into different clusters [64]. In the case of an abstract topological space, similar to the space constructed using the CED for semantic mobility sequences, the dendrogram used to visualize the results of hierarchical clustering provides support for understanding the studied space. However, in addition to defining a similarity measure, hierarchical clustering requires three other parameters to be defined: the strategy (top-down or bottom-up), linkage criterion, and dendrogram cutoff method.

Regarding the choice of strategy, the bottom-up approach has a polynomial time complexity of  $\mathcal{O}(n^2 \log n)$  versus an exponential complexity of  $\mathcal{O}(2^n)$  for the top-down approach [38]. Therefore, to handle a large dataset (in our case, 10 005 sequences), we used a hierarchical agglomerative clustering (HAC) algorithm based on a bottom-up strategy. A summary of hierarchical clustering algorithms in statistical software is presented in [69].

Regarding linkage criteria, the authors of [38] summarized the common options used in the literature. This choice depends on cluster shapes. The simplest approach, namely the single linkage criterion, is based on the minimum distance between a pair of elements from two clusters and can handle any cluster shape. However, repeated merges can lead to a chaining effect. In contrast, complete linkage, which is based on maximum distances, produces more compact clusters, but is sensitive to noise and outliers. Average linkage is particularly useful for convex clusters [38]. Because we do not know the shapes of clusters beforehand and we want robustness to outliers and immunity to chaining effects, we adopted the Ward criteria that minimize the total within-cluster variance. This is similar to the K-means algorithm, which is less affected by noise and tends to create convex compact clusters.

Finally, the determination of the optimal number of clusters can be considered from different perspectives and is a relatively difficult problem [33]. The simplest method for hierarchical clustering is based on higher relative loss of inertia criteria [41]. This method is associated with the largest gap between two successive agglomerations in a dendrogram. A summary of the different techniques can be found in [33].

Finally, quality clustering indicators such as the Silhouette score [65] are useful criteria for assessing the natural number of clusters and ensuring the validity of clustering. In particular, the Silhouette score is based on the same objective function as the Ward criterion and can be maximized to determine the optimal number of clusters.

## 3.5 Analysis of semantic sequence clusters

Let  $\{C_1, \dots, C_m\}$  be a partition of the dataset  $\mathcal{D}$  of semantic mobility sequences where  $C_{k \in [1, m]}$  represents a cluster. In this section of the pipeline, we wish to extract the meaningful characteristics of each cluster  $C_k$  in order to understand and explain the mobility behavior of each cluster. To this end, we calculate most of the indicators defined in Section 3.3 for each cluster  $C_k$ .

For numerical frequency distribution indicators, such as Indicator 1 or Indicator 7, we use boxplots to summarize and compare the distributions of each cluster. In contrast, for categorical frequency distribution indicators such as Indicator 2 and Indicator 4, according to the process described in [21], we use contingency tables, mosaic plots [24] and stack plots to visualize information. In this phase, the indicators are enriched with significance tests such as chi-squared test and Pearson residuals [32] in order to identify under- or over-representation of some variables, patterns or activities in the clusters. Cramér's  $V$  score is used to evaluate the strength of relationships between these variables and clusters.

**Indicator 8 (Pearson residuals)** Consider a sample of size  $N$  of the simultaneously distributed variables  $A$  and  $B$  with  $a_1, \dots, a_p$  and  $b_1, \dots, b_q$ , let  $(n_{ij})$ ,  $1 \leq i \leq p$ ,  $1 \leq j \leq q$  be the number of times the values  $a_i$  and  $b_j$  are observed, and let  $(n_{ij}^*) = \frac{n_{ij} \times n_{i+} \times n_{j+}}{N}$  be the theoretical values where:

- $n_{i+} = \sum_{j=1}^q n_{ij}$ , represents the column marginal for that cell,
- $n_{i+} = \sum_{j=1}^q n_{ij}$  represents the row marginal for that cell.

Then, the Pearson residuals  $r_{ij}$  [32] are defined as:

$$r_{ij} = \frac{n_{ij} - n_{ij}^*}{\sqrt{n_{ij}^*}} \quad (8)$$

Pearson residuals represent the strength and direction of the association between  $a_i$  and  $b_j$ . The strength is defined by the absolute value of the residual and the direction by its sign. Units are in standard deviations, meaning a residual greater than 2 or less than -2 represents a significant departure from the independence at the 95% confidence level.

By calculating Pearson residuals, we can determine how much the observed values deviate from the values in the case of complete independence. For example, an interesting subject is the departure of the frequency values of an activity  $x_i \in \Sigma$  in a given cluster  $C_j$ . If  $|r_{ij}| \geq 2$ , we can conclude that  $x_i$  has a statistically significant association with cluster  $C_j$ , where the sign indicates if  $x_i$  is under- (negative sign) or over- (positive sign) represented in  $C_j$ . However, statistical significance does not necessarily imply a strong association. There is a more standardized strength test called the chi-squared test. Statistical strength tests represent correlation measures. For the chi-squared test, the most commonly used measure is Cramér's V score [17].  $V$  varies from zero (corresponding to no association between variables) to one (complete association) and can reach one only when each variable is completely determined by the other.

Therefore, these measures can be used to characterize the activities or daily patterns in a cluster and provide partial information regarding the behaviors represented by patterns. However, these significance tests do not provide meaningful information regarding the order in which activities are conducted. The origin-destination matrix provides some additional information regarding the other of activities, but it cannot represent complete coherent behaviors in a cluster. However, indicators of centrality such as the medoid and the mode of a cluster can be used to extract an archetypal mobility sequence from the cluster.

**Indicator 9 (Mode)** Given a set of elements  $C$  and a similarity measure  $d : C \times C \rightarrow \mathbb{R}^+$ , the mode  $M$  of  $C$  is defined such that:

$$M = \underset{X \in C}{\operatorname{argmax}} \{f(X)\} \quad (9)$$

where  $f$  denotes the frequency function. Intuitively,  $M$  is the element which is the most-frequent in  $C$ .

**Indicator 10 (Medoid)** Given a set of elements  $C$  and a similarity measure  $d : C \times C \rightarrow \mathbb{R}^+$ , the medoid  $m$  of  $C$  is defined such that:

$$m = \underset{X \in C}{\operatorname{argmin}} \left\{ \sum_{Y \in C} d(X, Y) \right\} \quad (10)$$

Intuitively,  $m$  is the element that minimizes the distance to all other elements in  $C$ .

To validate whether the medoid  $m$  actually represents the elements of a cluster, it is essential to study the topology of the cluster  $C$ . Here,  $m$  is a good representative of  $C$  if the formed cluster is hyperspherical (i.e., the distribution of distances  $d(x, m)$  follows a power law which indicating that most of elements are near  $m$ ). Furthermore, hierarchical clustering achieves a complete partitioning of a dataset. Therefore, this analysis can identify outlier elements in clusters which can be considerate as the 5% of elements far away from the medoid<sup>2</sup>. Another measure for studying scattering and outliers that ignores the topology of clusters is cluster diameter.

**Indicator 11 (Diameter and Radius)** Given a set of elements  $C$  and a similarity measure  $d : C \times C \rightarrow \mathbb{R}^+$ , the diameter  $\text{diam}$  of  $C$  is defined as:

$$\text{diam}(C) = \max_{X, Y \in C} \{d(X, Y)\} \quad (11)$$

where  $\text{diam}$  represents the greatest distance between any pair of elements in the cluster. It should be noted that  $\text{diam}$  can also represent the most-distant pair of elements if  $\max$  is replaced with  $\operatorname{argmax}$  in Equation 11. Similarly, the radius  $\text{rad}$  of  $C$  is defined as:

$$\text{rad}(C) = \max_{X \in C} \{d(m, X)\} \quad (12)$$

where  $m$  is the medoid of  $C$  such as defined by Indicator 10.

Finally, analysis can be completed by calculating the Silhouette score of a cluster.

---

<sup>2</sup>Under the hypothesis of hyperspherical clusters

**Indicator 12 (Silhouette)** Let  $\{C_1, \dots, C_m\}$  a partition of the dataset  $\mathcal{D}$ . The Silhouette score [65] is a value which quantifies how is appropriately  $X \in C_k$  is clustered. It is defined as:

$$sil(X) = \frac{b(X) - a(X)}{\max\{a(X), b(X)\}} \quad (13)$$

where:

- $a(X) = \frac{1}{|C_k|-1} \sum_{Y \in C_k, Y \neq X} d(X, Y)$

- $b(X) = \min_{C_i \neq C_k} \frac{1}{|C_i|} \sum_{Y \in C_i} d(X, Y)$

Here,  $a(X)$  is the mean of the distance between  $X$  and all other elements in  $C_k$ . Therefore, it can be interpreted as a measure of how well  $X$  is assigned to its  $C_k$ . On the other hand,  $b(X)$  is the smallest mean distance from  $X$  to every points in the other clusters. The cluster with the smallest mean dissimilarity is said to be the “neighboring cluster” of  $X$  because it is the next-best fit cluster for point  $X$ .

Thus, the Silhouette score of a cluster  $C_i$  is defined by the arithmetical mean of  $sil(X)$  for each  $X \in C_k$ :

$$Sil(C_k) = \frac{1}{|C_k|} \sum_{X \in C_k} sil(X) \quad (14)$$

The next section illustrates the application of the proposed methodology to a real-world dataset and describes our findings in terms of mobility behaviors.

## 4 Case study

To test the proposed methodology, we used a set of real mobility sequences obtained from a French household travel survey called EMD.<sup>3</sup>. The goal of the EMD survey is to provide a snapshot of the trips undertaken by residents of a given metropolitan area, which can aid in understanding mobility behaviors and measure changes over time.

In this section, we describe the EMD data in terms of quality, semantics, and size. The dataset is complemented by a domain ontology describing activity semantics (step (a) in the methodology in Fig. 1). A statistical study and overview analysis of the dataset conclude this section (step (b) in Fig. 1).

### 4.1 EMD Rennes 2018 dataset

The studied dataset is called “EMD Rennes 2018” and it represents a household travel survey conducted in Rennes city and the surrounding area (Britany region of France). The survey was conducted from January to April of 2018 during weekdays. The data represent 11 000 people (at least five years old) from 8 000 households. This sample is considered to be statistically representative of 500 000 households and one million residents. The details of the data collection methodology and its quality are discussed in [14], and a summary of the results of the EMD Rennes 2018 survey is presented in [6]<sup>4</sup>.

The dataset consists of a set of mobility sequences, each of which represents the activities performed by one person over 24h. Table 6 lists the different activity labels used in the EMD mobility sequences. Two main classes are represented: stop activities and move activities. The former corresponds to daily static activities such as “staying at home”, “working” and “shopping”. The latter represents transportation activities such as “walking” or “driving a car.”

Mobility sequences are defined based on the Stop-Move paradigm [58]. Each stop activity is followed by one (or more) move activities. Therefore, the time dimension is only considered in terms of the order of the activities, resulting in a *compositional approach* to mobility analysis.

**Example 1** Consider the following activities performed by Sam during a day: “Sam starts her day at home. Then, she walks to the bus station and takes the bus to work. She spends her work time at her office and then walks home.”

---

<sup>3</sup>from French “Enquête Ménages-Déplacements”

<sup>4</sup>References in french

Table 6: Description of activities in the EMD data

Color	Aggregated activity	Emoji	Activity label and description
<i>Stop activities</i>			
	Home	🏡	<b>1:</b> main home; <b>2:</b> second home, hotel
	Work	💼	<b>11:</b> work in official work place; <b>12:</b> work at home; <b>13:</b> work in another place; <b>43:</b> look for a job; <b>81:</b> do a work round
	Study	🏫	<b>21:</b> day nursery; <b>22:</b> study at school (primary); <b>23:</b> study at school (college); <b>24:</b> study at school (high school); <b>25:</b> study at school (university); <b>26:</b> study in another place (primary); <b>27:</b> study in another place (college); <b>28:</b> study in another place (high school); <b>29:</b> study in another place (university)
	Shopping	🛍️	<b>30:</b> visit a shop; <b>31:</b> visit a shopping center; <b>32:</b> shopping in mall; <b>33:</b> shopping in medium or little shops; <b>34:</b> do shopping in market place; <b>35:</b> drive-through shopping
	Personal Care	🏥	<b>41:</b> health care; <b>42:</b> administration step
	Leisure	🎬	<b>51:</b> sport, cultural or voluntary activity; <b>52:</b> go for a walk or window-shopping; <b>53:</b> go in restaurant; <b>54:</b> visit family or friend; <b>82:</b> do a shopping tour (more than 4 consecutive activity 30)
	Accompany	🕒	<b>61, 63:</b> go with someone; <b>62, 64:</b> pick-up someone; <b>71, 73:</b> drop-off someone to a transport mode; <b>72, 74:</b> collect someone to a transport mode
	Other	?	<b>91:</b> other (detail in notes)
<i>Move activities</i>			
	Smooth	🚶	<b>100:</b> walk; <b>110:</b> ride location bike; <b>111:</b> ride bike; <b>112:</b> bike passenger; <b>193:</b> roller, skateboard, scooter; <b>194:</b> wheelchair; <b>195:</b> small electric machines (electric scooter, segway, etc)
	Motorized	🚗	<b>113:</b> motor bike driver ( $< 50cm^3$ ); <b>114:</b> motor bike passenger ( $< 50cm^3$ ); <b>115:</b> motor bike driver ( $\geq 50cm^3$ ); <b>116:</b> motor bike passenger ( $\geq 50cm^3$ ); <b>121:</b> car driver; <b>122:</b> car passenger; <b>161:</b> taxi passenger; <b>171:</b> car transport (work); <b>181:</b> van or truck driver (for activity 81); <b>182:</b> van or truck driver (for activity 81);
	Public transportation	🚠	<b>131:</b> urban bus passenger; <b>133:</b> subway passenger; <b>138, 139:</b> other public transportation passenger; <b>141, 142:</b> local public transportation; <b>151:</b> train passenger
	Other mode	✈️	<b>191:</b> sea transport; <b>192:</b> airplane; <b>193:</b> other modes (agricultural equipment, quad bike, ect);

The mobility sequence  $S$ , which is represented below, corresponds to Sam's activities. By using activity codes in Table 6, we have  $S = \langle 1, 100, 131, 11, 100, 1 \rangle$ . Alternatively, by considering aggregated activities, represented by emojis, we obtain the following representation:  $S_{agg} = \langle \text{🏡}, \text{🚶}, \text{🚠}, \text{💼}, \text{🕒}, \text{🏡} \rangle$ .

□

Throughout this paper, we will use activity codes and emojis to represent sequences both in examples and when analyzing real sequences. Among the 11000 sequences in the dataset (corresponding to 11000 surveyed individuals), we filtered those containing no moves (corresponding to people that stayed at home the entire). This resulted in a final dataset of 10005 mobility sequences.

## 4.2 Ontology

The activity concepts detailed in Table 6 are also structured in a knowledge graph (or ontology), as shown in Fig. 2. This knowledge graph refers to Definition 2 and is a hybrid of the EMD meronomy and the Harmonised Time Use Surveys (HETUS) [20]. Each color corresponds to a meta-category representing *aggregated activities*. first-level nodes

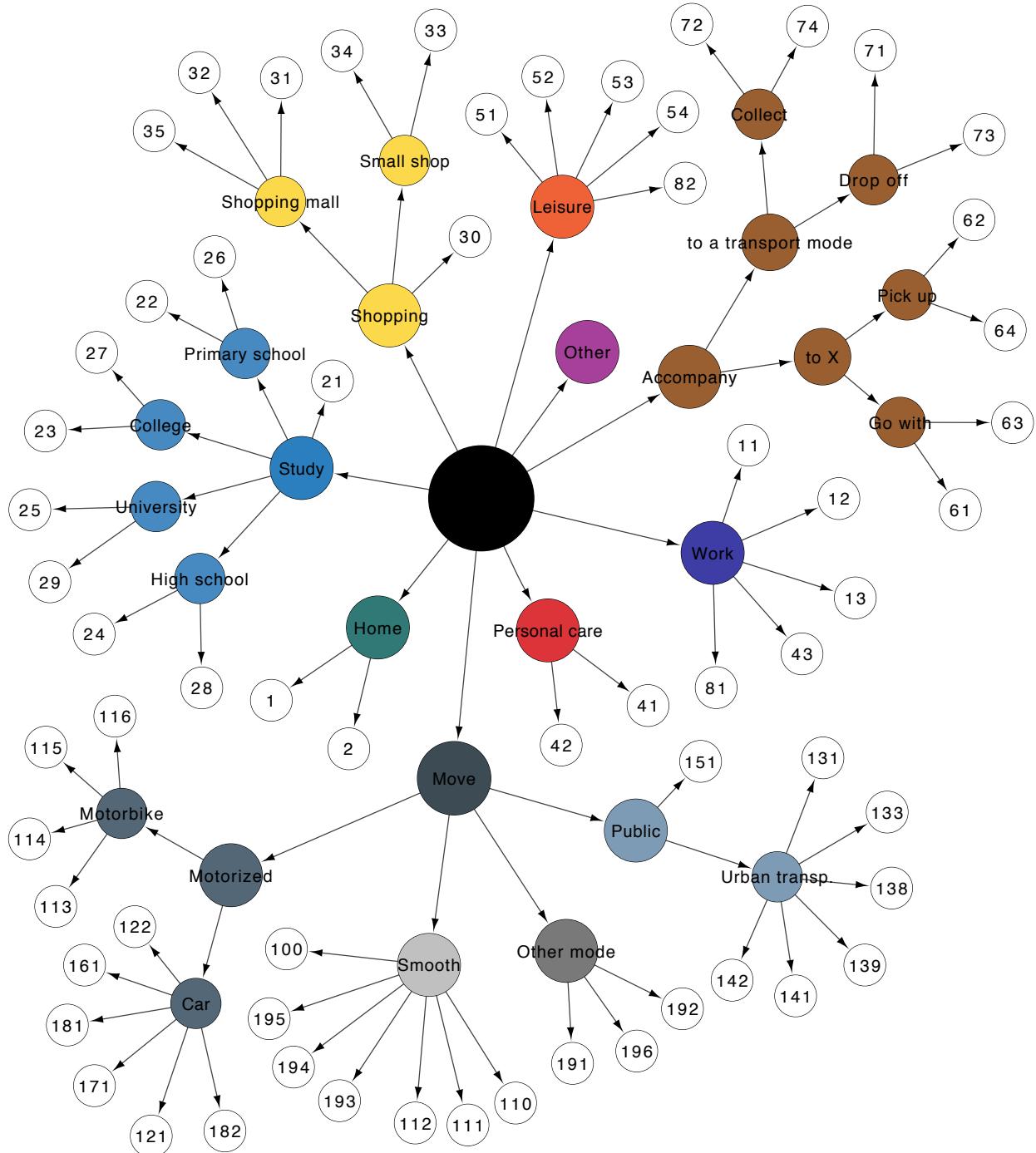


Figure 2: EMD graph ontology

for stop activities and second-level nodes for move activities (i.e., transport modes). Inter-level nodes come from the HETUS classification and first-level nodes and leaves come from the EMD survey.

Other possibilities for arranging concepts can be considered, each of which refers to a particular study context or specific business need. The structure of a graph influences the similarity measures between concepts.

**Example 2** Suppose we wish to compute the similarity between activities **100** (walking) and **121** (car driving). Using the ontology in Fig. 2, we can compute the Wu-Palmer similarity defined in Equation 1 as:

$$\begin{aligned}
sim_{WP}(100, 121) &= \frac{2 \times d(LCA(100, 121))}{d(100) + d(121)} \\
&= \frac{2 \times d(\text{Transport mode})}{d(100) + d(121)} \\
&= \frac{2}{7}
\end{aligned}$$

where  $LCA(x, y)$  is the Last Common Ancestor of concepts  $x$  and  $y$  and  $d(x)$  is the shortest path between node  $x$  and the root node (depicted in black in Fig. 2).

□

### 4.3 Comprehensive statistical analysis of the dataset

To understand the meaning of the data, we analyzed the entire dataset using the indicators described in Section 3.3 and summarized in 5.

Our first elementary study focused on the frequency of each activity in the sequences. For convenience, we separated the stop and move activities. Fig. 3 presents the distribution of each activity in the dataset. As predicted in [68], the frequency distribution follows a Zipf law. Intuitively, the three most frequent stop activities are 1 (home), 10 (work), and 33 (shopping in medium and small shops). For move activities, the most frequent items are 121 (car driving), 100 (walking), and 122 (car riding). This figure highlights the main activities that comprise the sequences.

We also performed a complementary study on the number of activities performed per day by an individual. Based on the stop-move representation, there are very few even sequence lengths. To overcome this issue, we consider intervals of length  $I_k$ . Fig. 4 presents the distribution of the lengths of the mobility sequences in the dataset. The green curve represents the estimated probability mass function of a Poisson distribution with a parameter  $\lambda$  obtained from maximum-likelihood estimation ( $\lambda = 1.36$ ). One can see that the intervals of lengths fit the Poisson distribution.

Another method for semantic sequence analysis is to study the transitions between symbols using an origin-destination matrix. Fig. 5 presents the transitions between two consecutive stop activities in the dataset. The ontology allows us to visualize these flows according to different levels of granularity. Detailed activities are presented on the left and aggregated activities are presented on the right. One can see that the home activity (🏠) plays a major role for most transitions, where 🏠 →  $x$  and the reverse  $x$  → 🏠.

In the daily mobility context, transitions were also studied in terms of individual mobility networks to identify topological patterns. Based on the work by Schneider et al. [66], we extracted the main motifs from the sequences. As shown in Fig. 6, the extracted motifs and frequencies are consistent with the results presented in [66]. We show the three most frequent motifs for 3, 4 and 5 nodes. We present the three most-frequent motifs for groups of three, four, and five nodes. Globally, one can see that the most-frequent patterns have less than four nodes and exhibit oscillations (labels I and III). Approximately 87% of the sequences follow one of the 11 identified motifs. Additionally, this analysis demonstrates that mobility sequences contain many stop activity repetitions.

Another technique for studying the repetition and regularity of a sequence  $S$  is to calculate the number of unique symbols  $\delta$  it contains. Fig. 7 presents the correlation between the length of a sequence  $|S|$  and the distinct number of activities  $\delta$ . The horizontal axis represents the interval length defined in Fig. 4 and the vertical axis represents the numbers of distinct moves  $\delta_{move}$  (left side) and number of distinct activities (stops + moves)  $\delta$  (right side). One can see that  $\delta_{move}$  remains globally stable with one or two different modes for any length of sequence. Therefore, we know that the diversity in the sequences stems from stop activities, while move activities are more often repeated. Regardless, according to the red curve in 7(b), one can see that most activities are repeated in a sequence.

Finally, the entropy and predictability of the mobility sequences can be studied to determine how sequences can be predicted. Fig. 8 portrays the distributions of these two variables. According to the number of activities in the sequences, the results are similar to those given by [68] and exhibit a low real uncertainty regarding a typical individual's location  $2^{0.4} \approx 1.32$ , which is less than two activities). It should be noted that these results are consistent with those presented Fig. 7 for the  $\delta$  values. The predictability in the random case is  $\Pi^{rand} \approx 0.24$ . One can see that the median number of different concepts in a sequence is four. This means that we can typically predict one out of the four previous activities. Unlike Song et al.'s results, our  $P(\Pi^{unc})$  distribution peaks approximately at  $\Pi^{unc} \approx 0.78$  which is similarly to the  $\Pi^{rand}$  value. This finding can be explained first by the small number of distinct activities in the sequences and also by the relatively small number of concepts in the dataset and the Zipf laws they follow (Fig. 3). This allow us to predict certain key activities (e.g., home, car, working, walking) based on the user activity history. Finally, the real predictability

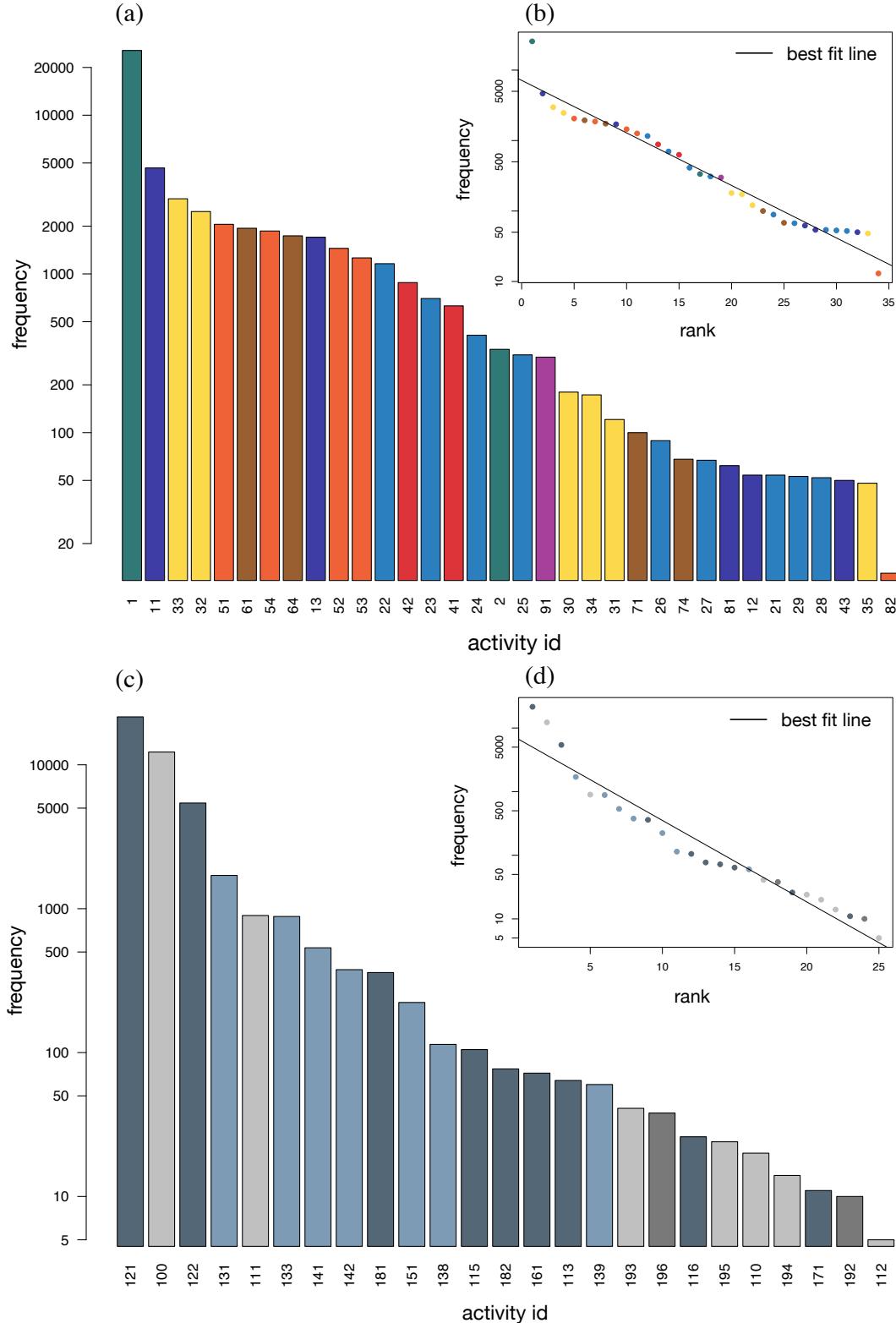


Figure 3: Stop (a) and move (c) activity distribution plot (a) log-plot showing the frequency of each stop activity codes, colors refer to aggregated activity. (b) and (d) show compatibility to a Zipf law model, each point correspond to activities in bar plot below.

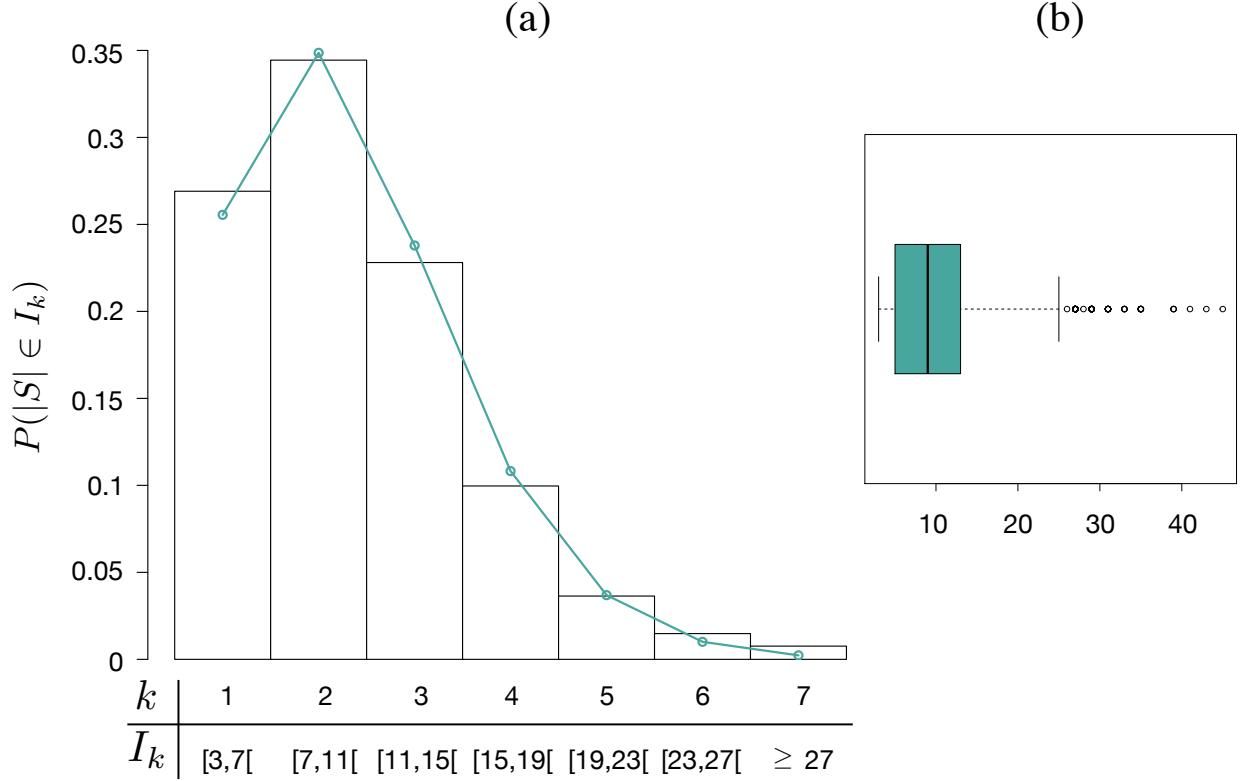


Figure 4: Length statistics of mobility sequences (a) The distribution of length  $|S|$  for a given interval  $I_k \in \{1\dots7\}$  follows a Poisson distribution  $P(|S| \in I_k) \approx \frac{1.36^k e^{-1.36}}{k!}$  (b) Box plot of the lengths

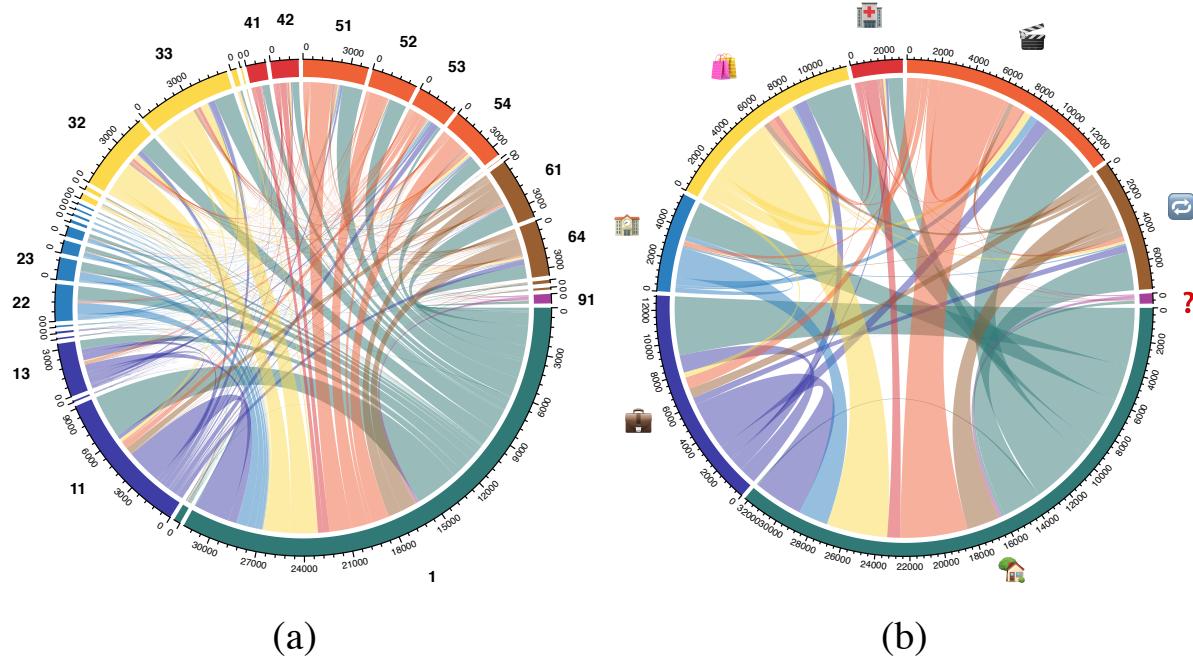


Figure 5: Chord diagram of flows between two consecutive stop activities (a) with all activities (b) with aggregated activities

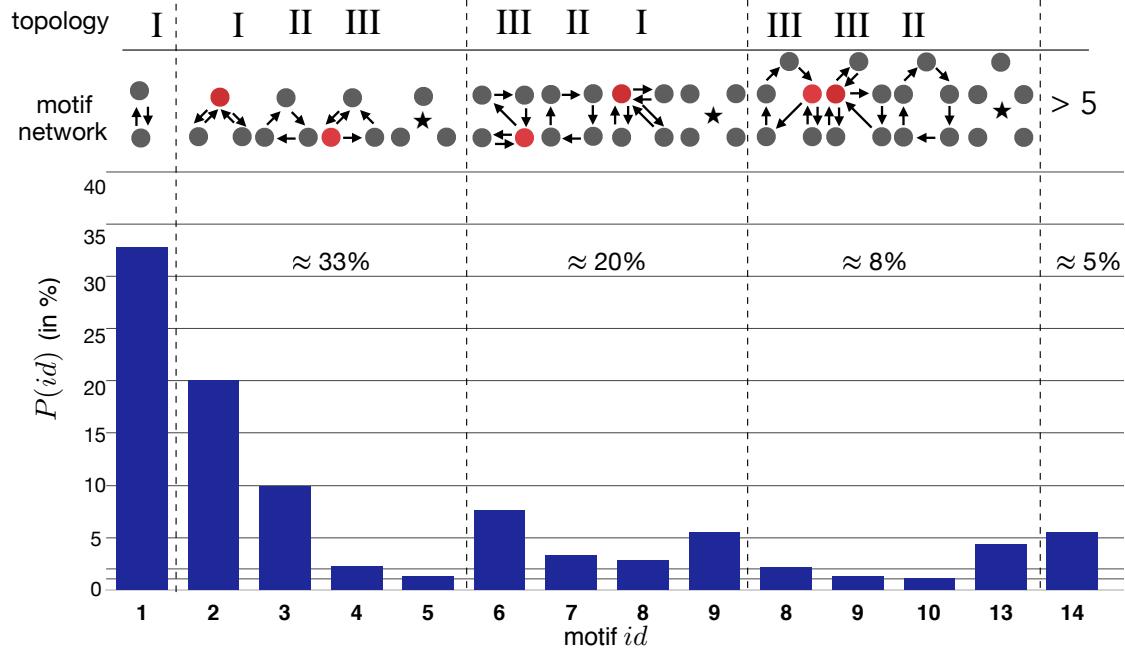


Figure 6: Daily mobility patterns. The motifs are grouped according to their size (separated by dashed lines).  $\star$  motifs include all other motifs with  $k \in \{3, 4, 5\}$  nodes. For each group, we show the estimated probability that a given motif has  $k$  nodes. The central nodes are highlighted in red. Motifs are classified by three rules indicating topological properties: (I) graphs with oscillations between two nodes, (II) graphs with cycles of 3 or more nodes and (III) graphs which combining both previous properties (I) and (II).

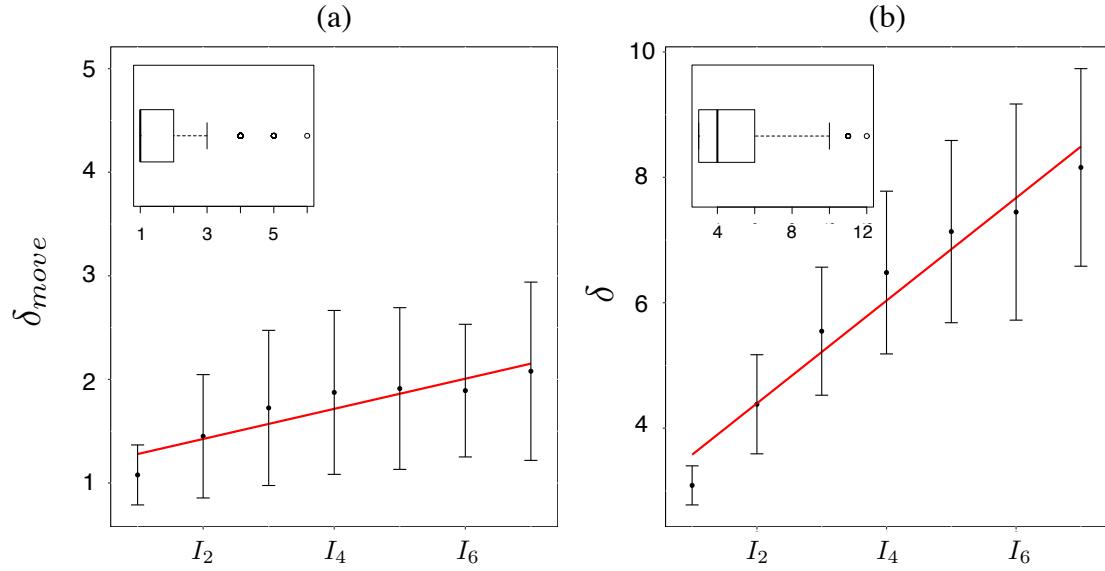


Figure 7: Correlation plots between intervals of length  $I_k$  and the number of (a) distinct move activities  $\delta_{move}$ , (b) distinct move + stop activities  $\delta$  in sequences. Box plot is showed for  $\delta$  and  $\delta_{move}$ . The coefficient of correlation is respectively (a)  $\rho = 0.4$  and (b)  $\rho = 0.8$ .

$P(\text{II})$  is peaked near  $\text{II}^{max} \approx 0.95$ , indicating that having a historical record of the mobility of an individual yields a high degree of predictability.

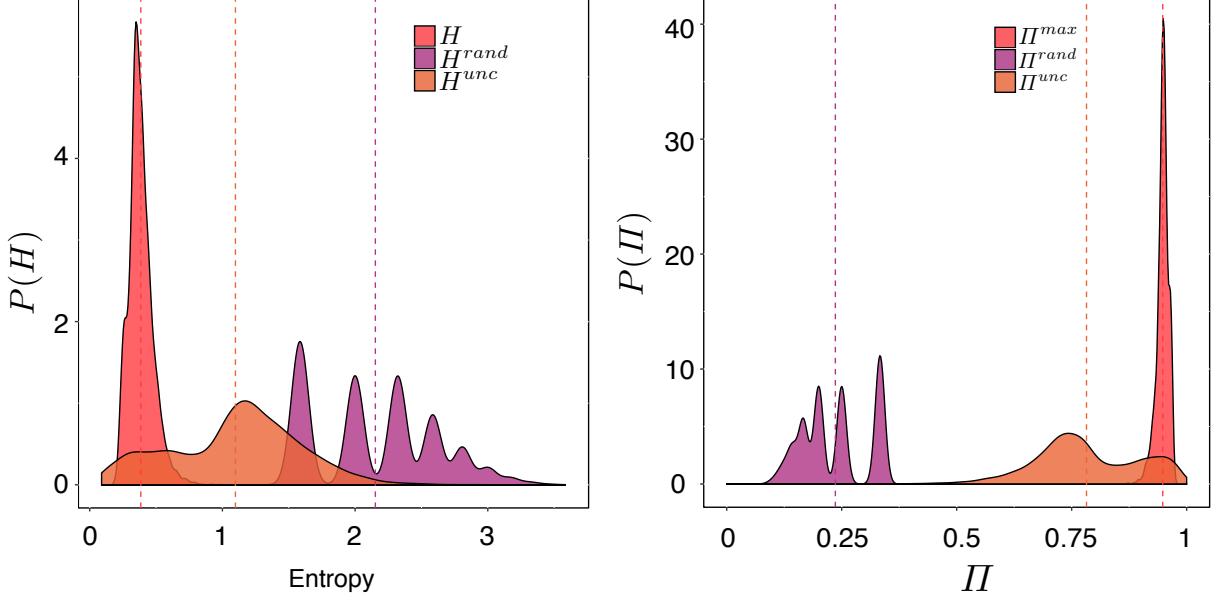


Figure 8: Entropy and predictability of the sequences, dash lines represent the mean (a) probability density function of the entropy  $H$ , the random entropy  $H^{rand}$ , and the uncorrelated entropy  $H^{unc}$  (b) Probability density function of the  $\Pi^{max}$ ,  $\Pi^{rand}$ , and  $\Pi^{unc}$

## 5 Semantic clustering behavior

This section describes the application of the steps (c) and (d) presented in the pipeline Fig. 1 applied to the EMD Rennes 2018 dataset. The first subsection describes the clustering process using the CED similarity measure and the HAC clustering algorithm Agnes [38] with R software. We cluster the individual semantic mobility sequences and analyse variations in daily activity types. In the second subsection, we extract typical behaviors from clusters by summarizing main characteristics and distinct patterns in terms of the indicators discussed in Sections 3.3 and 3.5. A discussion of the obtained results and alternatives methods concludes this section.

### 5.1 Clustering process

As discussed in Section 3.5, the clustering process is performed based on the CED measure and a hierarchical clustering algorithm. Here, we discuss the settings for these two methods and the validity of the clusters obtained in terms of quality scores.

#### 5.1.1 Similarity measure and HAC initialisation

As described in Section 3.4.1, the CED similarity measure requires the setting of several parameters. Empirically, we applied the following settings for CED during the clustering process:

- The  $\alpha$  coefficient is set to zero to give fully priority to context.
- The contextual vector was encoded using the Gaussian kernel below.

$$f_k(i) = \exp\left(-\frac{1}{2}\left(\frac{i-k}{\sigma}\right)^2\right)$$

where  $\sigma$  is a coefficient that controls the flatness of the curve around the activity at position  $k$ . The larger is  $\sigma$ , the more context surrounding the index  $k$  is considered. In our experiments, we used the value of  $\sigma = \frac{m}{2}$  where  $m$  is the median sequence size ( $m = 9$  according to Fig. 4). Therefore,  $v_i(e) = f_k(i)$ .

With these settings, the CED is a semi-metric, meaning it satisfies the requirements of symmetry and identity of indiscernible, but the triangle inequality does not hold.

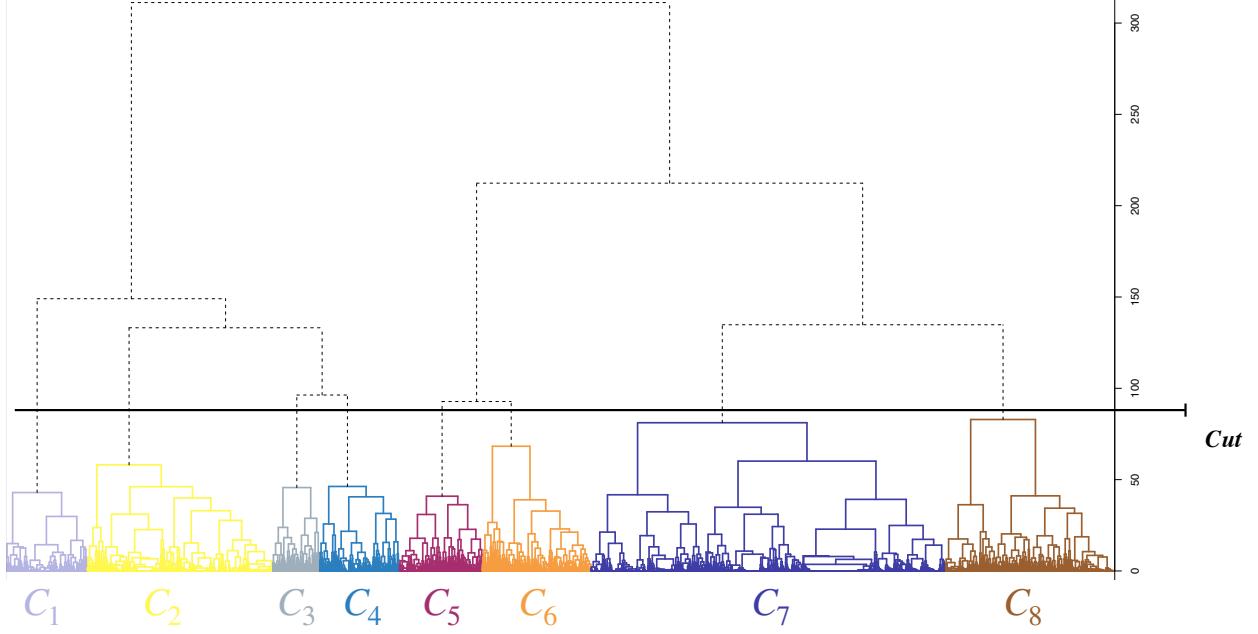


Figure 9: Dendrogram of the HAC clustering algorithm of the EMD 2018 dataset. Eight clusters are formed by the cut of the dendrogram.

Table 7: Cardinal number, Silhouette indice and diameter and radius of each cluster

Cluster $C_i$	$ C_i $	% (in total)	$Sil(C_i)$	$diam(C_i)$	$diam(C_i^{95\%})$	$rad(C_i)$	$rad(C_i^{95\%})$
1	738	7.4	0.41	8.85	5	5.04	3.44
2	1673	16.7	0.37	20.03	8	12.66	4.68
3	423	4.2	0.01	20.81	7.74	7.95	5.53
4	719	7.2	0.12	26.64	7.21	12.51	5.7
5	747	7.5	0.18	23.42	6.9	9.86	5.35
8	981	9.8	0.1	24.34	8.15	14.59	6.11
7	3199	32	0.29	20	5.57	11.09	4.09
12	1525	15.2	0.07	28.5	7	10.14	4.65

Regarding the HAC algorithm, because we do not know the shapes of the clusters, but we want to preserve robustness to outliers and immunity to chain effects, we propose using the Ward criteria, which minimizes the total within-cluster variance, leading to the generation of convex compact clusters that are less affected by noise.

### 5.1.2 Clustering validity

One problem in an unsupervised clustering process is to determine the optimal number of clusters that best fits the inherent partitioning of the dataset. In other words, we must evaluate the clustering results for different cluster numbers, which is the main problem in determining cluster validity [33]. There are three main approaches to validating clustering results: (1) external criteria, (2) internal criteria, and (3) relative criteria. Various indices are available for each criterion.

The structure of HAC and the formed clusters are presented in Fig. 9. In our study, because we did not have a predetermined cluster structure, we used internal validation indices, whose fundamental goal is to search for clusters whose members are close to each other and far from the members of other clusters. Specifically, we used two indices to select the optimal number of clusters. The first is the *inertia gap*, which represents the total distance between two consecutive agglomerations. The wider is the gap, the greater the change in cluster structure and the greater the Silhouette index, which reflects the compactness and separation of clusters. The *Silhouette index* is defined in the range  $[-1, 1]$ . A higher value Silhouette index indicates a better clustering result.

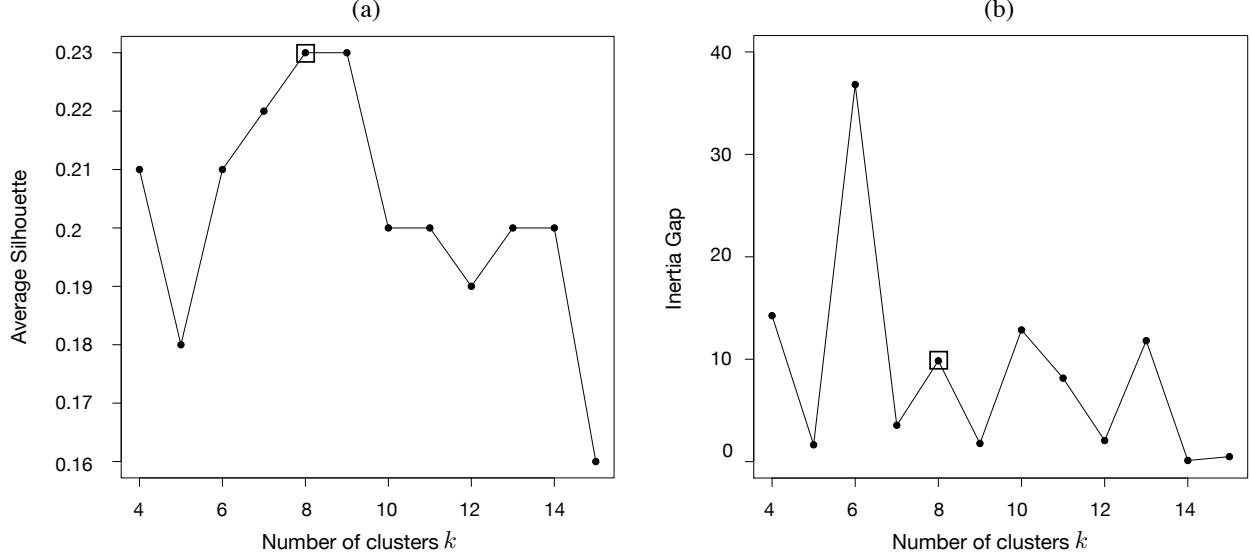


Figure 10: Clustering validity indices (a) Average Silhouette (b) Inertia gap

Fig. 10 presents graphs of (a) the average Silhouette index and (b) inertia gap. The relatively low Silhouette values can be attributed to the particular topology associated with CED combined with the Ward criterion and the presence of outliers.<sup>5</sup>. Because we want a number of clusters greater than five to ensure correct analysis, plot (a) suggests the choice of eight or nine clusters. Values of 6, 7, 10, 11, 13, or 14 could also be used. Plot (b) strongly encourages the choice of six clusters, but 8, 10, or 13 clusters could also be used. According to these results, we **set the number of clusters to eight** for further analysis. Regardless, the choice of six clusters for narrow analysis, or 10 or 13 clusters for wide analysis may be feasible.

Additional information regarding the clusters, such as proportions, Silhouette index, diameters and radii are given in Table 7.  $C_i^{95\%}$  indicates that we filtered 5% of the most extreme values from the distribution. Therefore, the difference between  $diam(C_i)$  and the diameter of 95% of the elements in  $C_i$ , denoted as  $C_i^{95\%}$ , indicates that there is a proportion of outliers far away from the other elements in the cluster  $C_i$ . Similar radii values of  $C_i^{95\%}$  support this analysis.

## 5.2 Behavior extraction and cluster explanation

In this section, we reuse the indicators and statistics presented in Table 5 and Section 4, but enhanced with significance tests, to infer typical behaviors from clusters discovered in Section 5.1.2. This can help us to check the validity and interpretability of their patterns.

First, we analyse the lengths of the sequences inside the clusters. Fig. 11 presents the box plots for the sequence lengths in each cluster. Compared to the distribution of lengths and the box plot for the entire dataset (leftmost plot), one can see that clusters  $C_1$ ,  $C_2$  and  $C_7$  contain relatively short sequences with a median lengths of six and seven activities, corresponding to intervals  $I_1$  and  $I_2$  in the Poisson distribution (see Fig. 4). In contrast, clusters  $C_5$ ,  $C_6$  and  $C_8$  contain longer mobility sequences but have large length dispersions. Analogously, clusters  $C_3$  and  $C_4$  have middling sequence lengths corresponding to intervals  $I_2$  and  $I_3$ . The overlapping of box plots (i.e., the existence of several clusters containing sequences of the same length) and the distribution of outliers in the clusters indicate that sequence length as not a major criteria for grouping sequences. We claim that is an advantage of CED w.r.t other OM similarity measures.

Regarding the distributions of activities, Fig. 12 portrays the proportions of aggregated activities<sup>6</sup>. An interesting effect that can be observed in this graph is the strong discrimination and stratification effect of clusters according to move activities. Motorized transport activities are very common in clusters  $C_7$  and  $C_8$ , but other move activities are not. Clusters  $C_2$  to  $C_6$  stand out based on their large proportion of smooth move activities whereas  $C_1$  and  $C_3$  contain

<sup>5</sup>Note that Silhouette is particularly suitable for hyper-spherical clusters like the one constructed by K-means algorithms.

<sup>6</sup>Thanks to the ontology, we can select the level of granularity of our analysis. Aggregated activities having been retained in order to avoid cognitive overload on graphs.

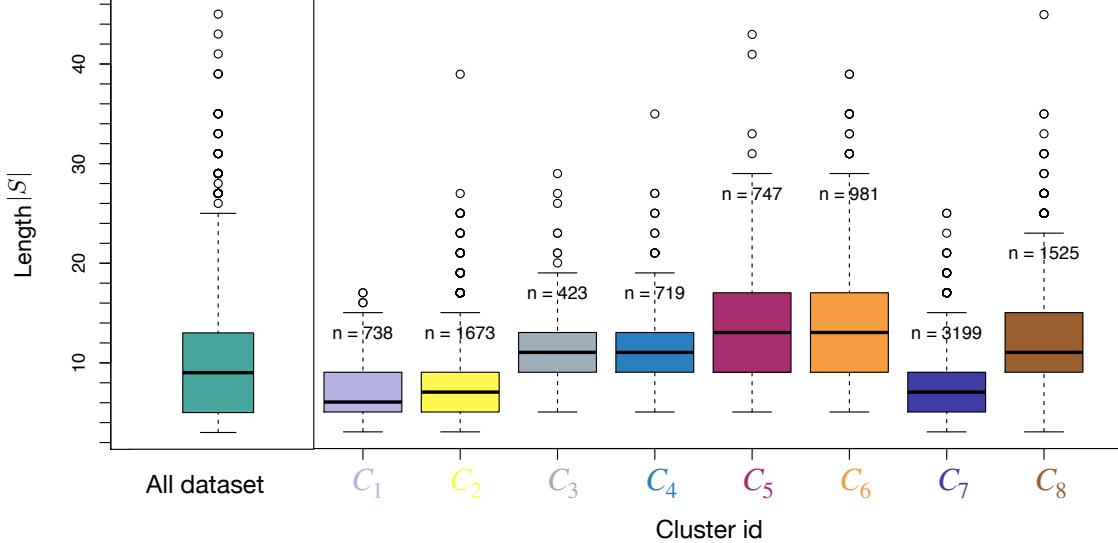


Figure 11: Box plots of sequences' length in each cluster

many public transportation move activities. Several stop activities are also a distinctive features in certain clusters. For example, school activities are particularly popular in clusters  $C_1$  and  $C_4$ , while work activities in clusters  $C_5$ ,  $C_7$  and  $C_8$  and accompany activities being common in cluster  $C_8$ . Similarly, some clusters tend to contain very few instances of certain activities. For example, cluster  $C_6$  contains very few work or study activities

To analyze theses over- and under-representations, we generated a mosaic plot combined with Pearson residuals to quantify the departure of each cell from independence. Fig. 13 presents the mosaic plot with residuals between clusters and aggregated activities.

We now recall several rules for the interpretation of this type of plot. Let each line represent a cluster an aggregated activity  $aggAct_i$  and each column represents  $C_j$ . We let  $c_{ij}$  denote the cell in line  $i$  and column  $j$ :

- The width of  $c_{ij}$  is proportional of the size of  $C_j$ .
- The height of  $c_{ij}$  is proportional the number of  $aggAct_i$  under the condition of to be in  $C_j$ .
- The area of  $c_{ij}$  is proportional to the frequency of  $aggAct_i$  and  $C_j$ .

The color of a cell  $c_{ij}$  indicates the value of the corresponding Pearson residual  $r_{ij}$ . A blue shaded cell indicates an over-representation of the aggregated activity  $aggAct_i$  in  $C_j$ . A red-shaded cell indicates an under-representation. Based on this graphical representation, it is easy to visualize the proportion of a given activity in each cluster. For example, one can immediately see that approximately 40% of cluster  $C_1$  is comprised of public transportation activities. Additionally, based on the residuals, we can immediately and easily identify the characteristic activities in a cluster, as well as those that are under-represented. Therefore, the Fig. 13 complements and validates our previous analysis based on a stacked plot (Fig. 12) based on quantitative Pearson residuals. The Cramér's  $V$  coefficient provides information regarding the associations between clusters and aggregated activities. The good value of  $V$  (0.3) highlights the quality of association between our clusters and the activities performed in mobility sequences. The low number of white cells in the mosaic plot confirms our choice of clustering process.

Regarding transitions between activities, Fig. 14 presents a chord diagram for each cluster. As a consequence of the analysis presented in Fig. 7 which indicated transport mode remains globally stable within sequences, we only represent stop activity transitions. Flows are represented between two leaf activities in the ontology to explore the content of clusters in detail. For example, one can see that cluster  $C_1$  contains study activities ranging from junior high school (23) to university (25), whereas cluster  $C_4$  mainly contains school children (22). An explanation for this split can be derived from the fact that cluster  $C_1$  mainly concentrates on public transportation activities (see Figs. 12 and 13), which is generally related to teenagers, whereas older students tend to be autonomous. However, younger children are mainly accompanied to school by their parents by car or on foot, which can be observed in cluster  $C_4$ . This analysis is supported by Table 8, where the centrality indicators highlight typical sequences. For example, the most frequent sequence in  $C_1$  (mode) involves traveling to and from school via public transportation. Cluster  $C_4$  mainly focuses on car and foot travel, but also includes some leisure activities (51, 53).

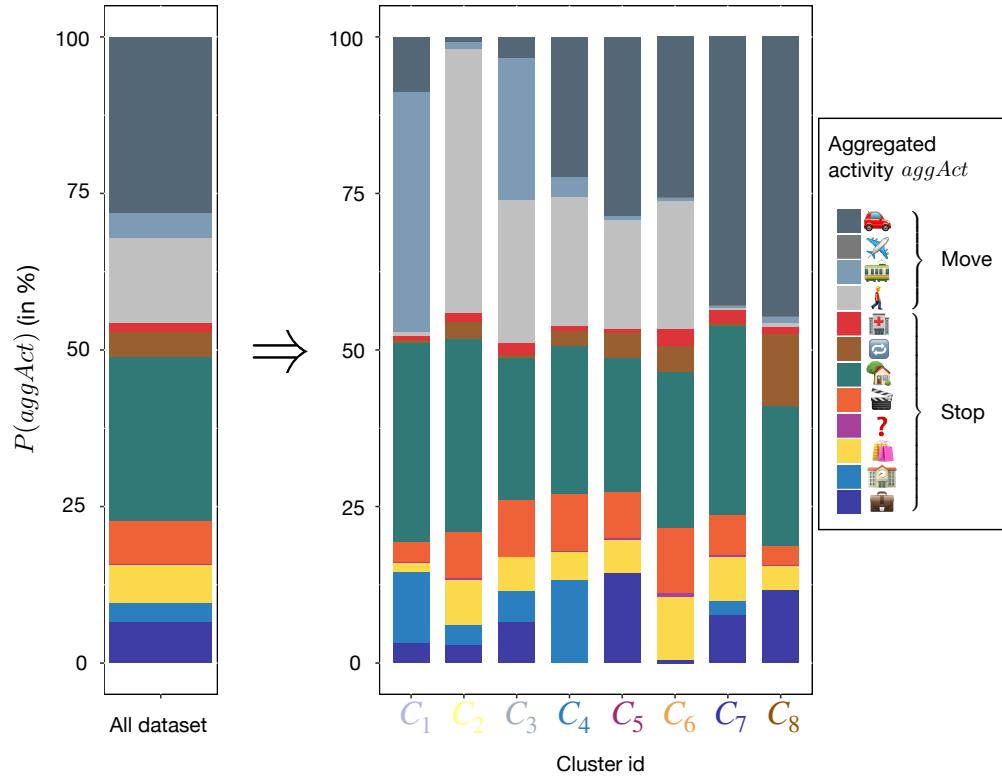


Figure 12: Stacked plot of the proportion of aggregated activities ( $aggAct$ ) in all dataset on the left and in each cluster on the right

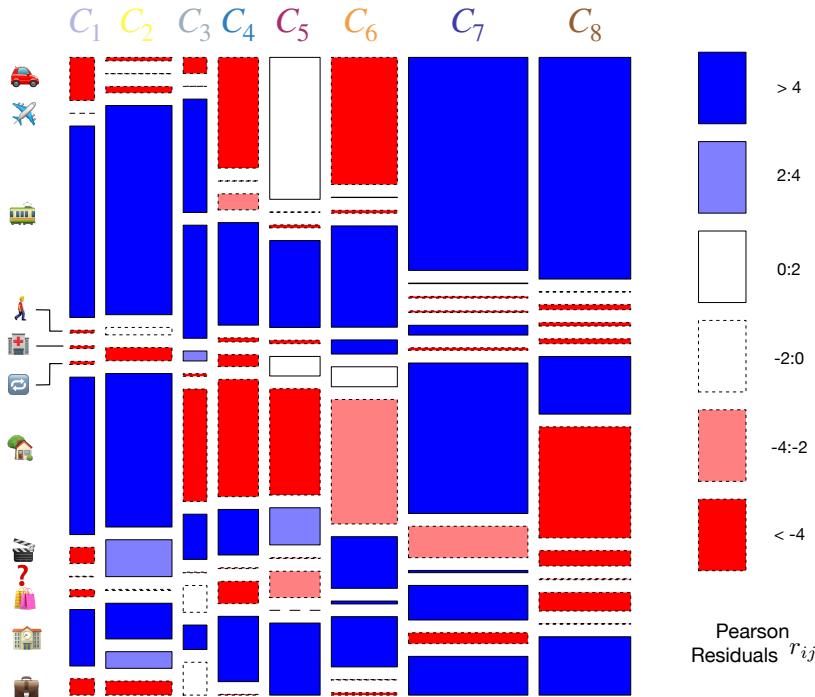


Figure 13: Mosaic plot and Pearson residuals between aggregated activities and clusters. Cramér's  $V = 0.3$

Table 8: Centrality indicators in each cluster

Cluster $C_i$	Medoid	Mode
1	$\langle 1, 131, 23, 122, 1 \rangle$	$\langle 1, 141, 23, 141, 1 \rangle$
2	$\langle 1, 100, 33, 100, 1 \rangle$	$\langle 1, 100, 33, 100, 1 \rangle$
3	$\langle 1, 131, 131, 11, 100, 53, 131, 1 \rangle$	$\langle 1, 141, 23, 100, 27, 100, 23, 141, 1 \rangle$
4	$\langle 1, 122, 22, 100, 51, 122, 1 \rangle$	$\langle 1, 122, 22, 100, 1 \rangle$
5	$\langle 1, 121, 11, 100, 53, 100, 11, 121, 1 \rangle$	$\langle 1, 121, 11, 100, 53, 100, 11, 121, 1 \rangle$
8	$\langle 1, 100, 33, 100, 1, 121, 52, 121, 1 \rangle$	$\langle 1, 121, 33, 121, 1, 100, 52, 100, 1 \rangle$
7	$\langle 1, 121, 11, 121, 1 \rangle$	$\langle 1, 121, 11, 121, 1 \rangle$
12	$\langle 1, 121, 61, 121, 11, 121, 64, 121, 1 \rangle$	$\langle 1, 121, 13, 121, 1 \rangle$

Regarding the worker cuters (i.e.,  $C_5$ ,  $C_7$  and  $C_8$ ), we observed different behaviors for each one. In cluster  $C_5$ , the typical behavior appears to be that of a worker driving to work (11, 13) and then walking to a restaurant for lunch (53) before returning to work and then driving home. This scenario is supported by the medoid mobility sequences and Fig. 15 which represents the daily patterns in each cluster. In  $C_5$ , one can see a trend of oscillation between two activities with a central node. There are also some activities that can be added to the semantic sequence, such as shopping (32, 33) after work, going for a walk or window-shopping (52), or accompanying activities (61, 64). Cluster  $C_7$  represents individuals who travel to an activity, typically work (11), by car, then return home by car again. This interpretation is consistent with the short semantic mobility lengths in the cluster and the large majority of daily patterns with a single oscillation. This mobility behavior is the most common in the dataset and can be interpreted as the daily routine of going to work by car, occasionally shopping in a mall (32), and then going back home. Cluster  $C_8$  is focused on workers that accompany and pick up (61, 64) someone before and after work (11, 13) with a possible mobility around the workplace (13). Similarly, cluster  $C_7$  moves are almost exclusively performed by car. Furthermore, in this cluster, mobility sequences are relatively long and form complex patterns with, generally, four or more stop activities.

Finally, some clusters can be distinguished by the absence of some common elements. For example, the people in cluster  $C_6$  do not work or study and they tend to spend their time mainly on shopping or leisure activities. Additionally, box plot in Fig. 11 reveals that the mobility sequences in  $C_6$  are long. Lastly, clusters  $C_2$  and  $C_3$  are especially characterized by their move activities. Individuals in  $C_2$  almost exclusively move on foot to perform a single activity. Compared to the mobility sequences in  $C_6$ , these sequences are relatively short and based on a single oscillation between home and another location. The typical behavior in  $C_3$  seems to be that people who use both public transportation and walking for mobility.

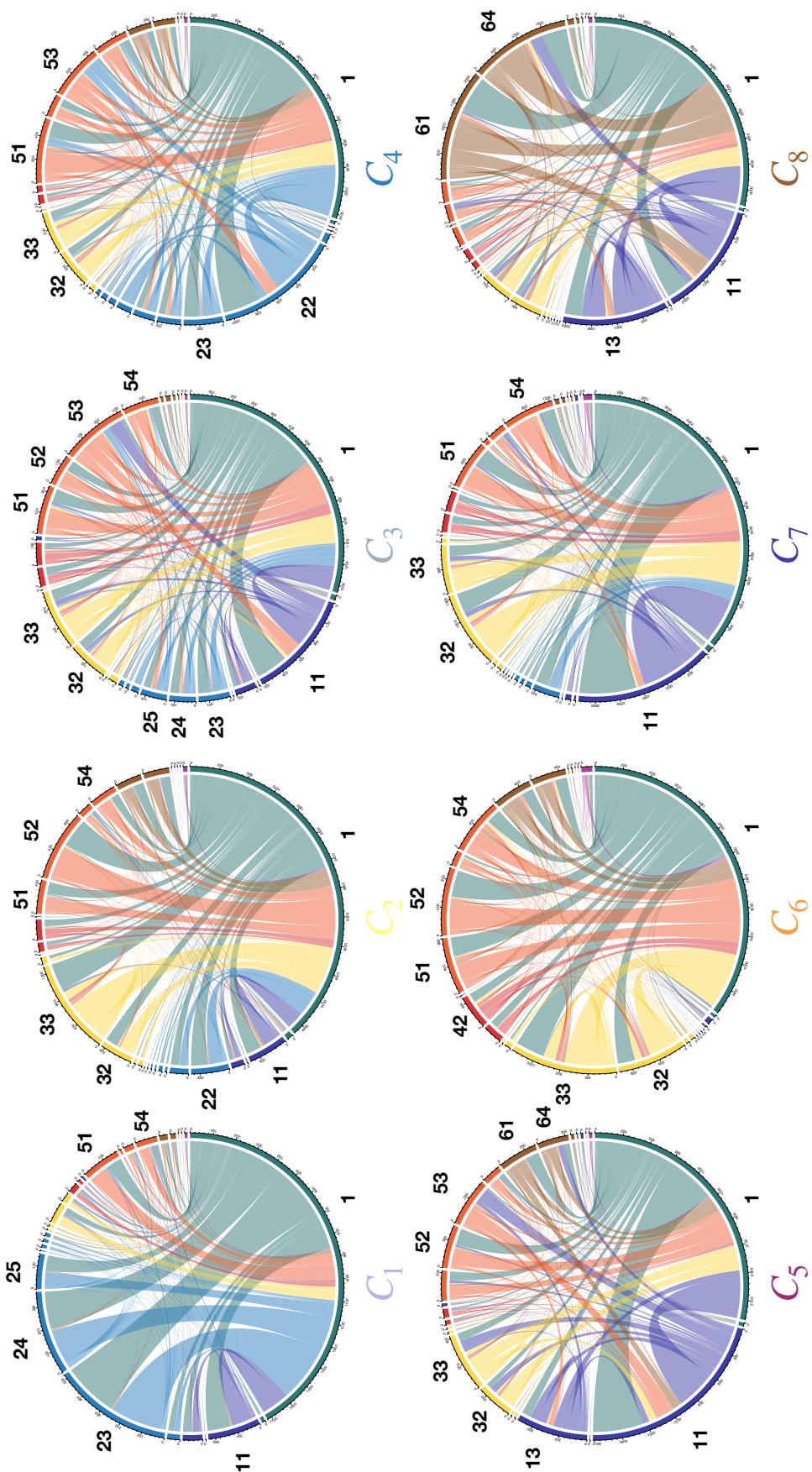


Figure 14: Chord diagrams of Stop activities in each cluster

	1	2	3	4	5	5	7	8	9	10	11	12	13	> 5	$\sum_j$
motif id	1	2	3	4	5	5	7	8	9	10	11	12	13	> 5	$\sum_j$
$C_1$	542	84	63	8	8	14	11	1	4	2	0	0	1	0	738
$C_2$	912	321	147	28	16	61	29	39	50	13	9	5	27	16	1673
$C_3$	15	168	51	10	6	48	13	14	35	8	1	9	28	17	423
$C_4$	65	219	40	18	6	89	34	34	56	29	19	17	44	49	719
$C_5$	7	213	12	9	11	95	6	33	108	38	16	5	81	113	747
$C_6$	38	312	27	28	13	110	14	85	60	33	40	8	88	125	981
$C_7$	1578	496	481	84	51	182	90	26	114	20	10	14	37	16	3199
$C_8$	147	211	179	46	26	165	142	54	134	73	35	56	131	126	1525
$\sum_i$	3304	2024	1000	231	137	764	339	286	561	216	130	114	437	462	10005

Figure 15: Heat map with Pearson residuals of daily patterns in each cluster. Cramér's  $V = 0.25$ 

### 5.3 Semantic mobility behavior discovering

Based on our previous analysis of clusters, we extract a global behavior from each cluster. Table 9 summarizes the eight discovered behaviors. The columns "Typical activities", "Length" and "Daily patterns" were computed using the Algorithm 2 and represent the predominant activities, median lengths of sequences (intervals), and the predominant daily patterns, respectively. For the sake of brevity, typical activities were extracted at the aggregated activities level (using emojis). Finally, the "Behavior" column contains, mnemonic labels which summarizes the analysis carried out Section 5.2.

We can summarize the behaviors in the clusters as follow: Cluster  $C_1$  contains a majority of short mobility sequences, with only one loop between home and middle/high school or university, and an extensive use of public transportation such as buses. This group mainly consists of *Teenagers* mobility behavior.

---

#### Algorithm 2: Behavior Discovery summary

---

**Data:** Set of clusters  $\mathcal{C} = \{C_1, \dots, C_k\}$

**Result:** Typical activities, Length and Daily patterns

**for**  $C_i \in \mathcal{C}$  **do**

▷ medoid( $C_i$ ) and mode( $C_i$ ) refer to Table 8.  $f(x, C_i) = \sum_{S \in C_i} \text{count}(x, S)$  denotes the frequency of activity

$x$  in all sequences of  $C_i$ . Typical activities( $C_i$ ) =  $\{x | x \in \Sigma \wedge \text{PearsonResiduals}(f(x, C_i)) \geq 4 \wedge (x_i \in \text{medoid}(C_i) \vee \text{mode}(C_i))\}$

▷  $I_k$  refers to intervals of Poisson distribution in Fig. 4

Length( $C_i$ ) =  $\begin{cases} \text{"Short"} & \text{if } \text{median}(\{|S| : S \in C_i\}) \in I_1 \\ \text{"Medium"} & \text{if } \text{median}(\{|S| : S \in C_i\}) \in I_2 \\ \text{"Long"} & \text{else} \end{cases}$

▷  $\mathcal{G}$  refers to a dictionary of networks from Algorithm. 1.

DailyPatterns( $C_i$ ) =  $\{G_S | S \in C_i, \text{PearsonResiduals}(\mathcal{G}[G_S]) \geq 4\}$

**end**

---

Table 9: Summary of discovered behaviors

Cluster $C_i$	% (in total)	Typical activities	Length	Daily Patterns (motif id)	<i>Behavior</i>
1	7.4	{bus, home}	Short	1	<b>Teenagers</b>
2	16.7	{walk, shop}	Short	1	<b>Foot shoppers</b>
3	4.2	{bus, walk, building, cinema}	Medium	2, 5	<b>Mixed transportation</b>
4	7.2	{walk, building, cinema}	Medium	2, 4	<b>Schoolchildren</b>
5	7.5	{walk, briefcase, cinema}	Long	2, 5, 9, 13, 14	<b>Wandering workers</b>
8	9.8	{walk, shop, cinema}	Long	2, 5, 8, 11, 13, 14	<b>Shopping addicts</b>
7	32	{car, briefcase}	Short	1, 3	<b>Daily routine</b>
12	15.2	{car, briefcase, laptop}	Medium	5, 7, 9, 10, 12, 13, 14	<b>Working parents</b>

Cluster  $C_2$  is characterized by people who only walk for shopping, which we call the *Foot shoppers*.

The main feature in  $C_3$  is that the sequences combine walking and public transportation, which we call them *Mixed transportation* people. *Schoolchildren* are mainly clustered in  $C_4$ , with a large proportion of primary school activities, followed by sports or cultural activities. These individuals mainly move by walking or by riding in cars.

In cluster  $C_5$ , the prototypical behavior is that of an individual working and going out for lunch, typically at restaurant. These *Wandering workers* achieved their mobility by driving between home and work and, walking between work and place for food/leisure.

The representative behavior of individuals in cluster  $C_6$  is that they do not work or study. They spent the majority of their time in shopping or leisure activities. We refer to these individuals as the *Shopping addicts*.

Cluster  $C_7$  is the largest cluster and contains 32% of the dataset. Individuals in  $C_7$  done mainly short mobility sequences that represent people who go to work by car and then travel back home. This behavior, with its elementary activities (car, work, and sometimes shopping at a mall) and oscillation patterns, evokes a simple *Daily routine*.

Finally,  $C_8$  represents a similar behavior to that of  $C_7$  but individuals typically transport somebody else by car before working and then pick them back up after work. This behavior can be interpreted as parents accompanying their children to school in the morning and picking them up in the evening. Therefore, we refer to these individuals as *Working Parents*.

Figures 16 presents a graphical summary of the clusters and corresponding behaviors. The area of each square is proportional to the size of the associated cluster. The colors and compositions refer to the dendrogram in Figure 9.

## 5.4 Discussion

In the previous subsection, we presented the analysis and results of the clustering process according to the methodology introduced in Section 3.5. This facilitated the discovery of several interesting and coherent patterns of semantic mobility, which are summarized in Table 9.

Regardless, several problems and alternatives should be considered. First of all, as discussed in Section 2, there are many different similarity measures for semantic sequences. The choice of a measure has a significant impact on the results of clustering. In this paper, we used CED, which is an alternative measure to those mentioned in Table 1, which could also be considered. The setting of CED: the similarity measure between activities, the ontology, the contextual vector and the  $\alpha$  coefficient are all parameters that can be modified to change the clustering results. We experimentally tuned each parameter and referred to business knowledge for the construction of our ontology.

The second point is the choice of the clustering algorithm. As indicated in Table 7, the diameters of the clusters indicate the presence of some outliers and the Silhouette scores suggest that the clusters are not hyper-spherical. Therefore, the use of a density-based clustering algorithm such as DBSCAN or OPTICS, combined with more complete study of the

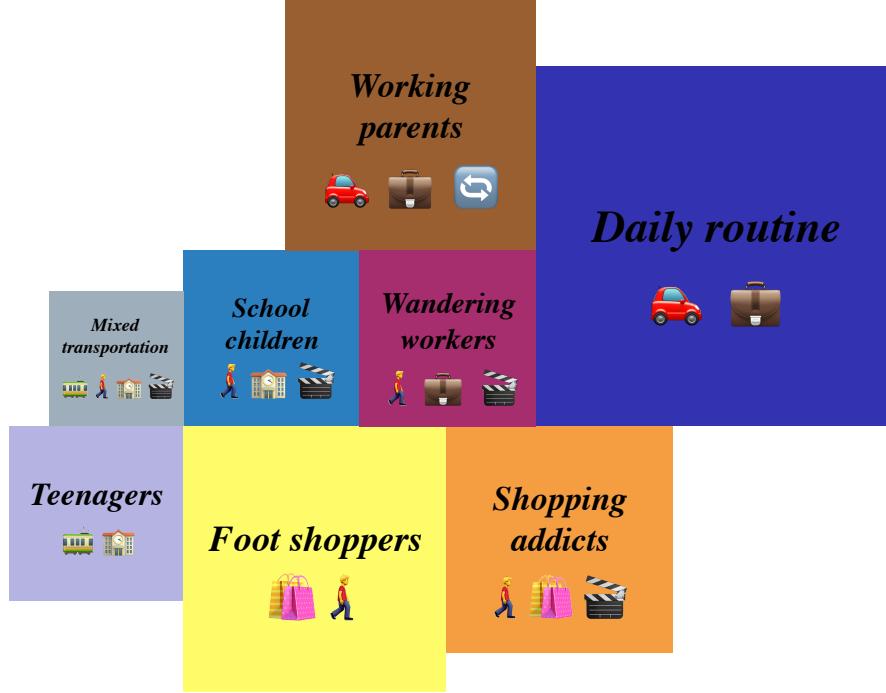


Figure 16: Graphical summary of discovered clusters

topological space and neighborhood relationship via UMAP, as well as intra- and inter-cluster distances, could help us to obtain denser clusters and detect outliers.

The final point is the level of analysis of the activities in the ontology. To prevent cognitive overload during visualization, Section 5.2 only presented the results of aggregated activities. Regardless, a detailed analysis at the level of leaf activities in the ontology would also be relevant and could refine the discovered behaviors.

Based on the proposed methodology, we were able to analyze and extract precise cluster behaviors from both socio-cognitive and urban perspectives. Therefore, this approach should be helpful for expert analysts in terms of limiting psychological biases, such as confirmation bias.

The proposed methodology supports the comprehension of clusters and is useful for the evaluation and tuning of clustering methods. The discovery of coherent and meaningful behaviors may trigger the proposal of novel metrics for the quality of experimental setups and relevance of various methods.

## 6 Conclusions and future work

In this paper, we introduced a novel methodology, called SIMBA, to mine, discover and analyze behaviors in semantic human mobility sequences. The proposed process is generic and can be adapted for any sequence of categorical data.

SIMBA introduces a simple and complete pipeline from raw data to clustering analysis for studying semantic mobility sequences and extracting mobility behaviors. SIMBA leverages the use of a hierarchical clustering algorithm combined with CED to cluster similar mobility sequences.

Based on an extended literature review of both human mobility properties and semantic similarity measures, we selected complementary statistical indicators to describe semantic mobility sequences from different points of view. To the best of our knowledge, SIMBA is the first complete and modular methodology supporting the understanding of human behaviors with a large panel of visual indicators that highlight the complementary properties of semantic mobility.

The proposed approach was tested on a real dataset of 10005 semantic mobility sequences from a household travel survey. We were able to identify specific behaviors that can constitute key information on urban activities. Thanks to the proposed methodology, discovered clusters are easily interpretable and sound coherent with our intuition. Furthermore, the clusters revealed regular patterns of human daily activities that are consistent with previous findings regarding

the strong predictability and regularity of human mobility. We hope that our methodology will be helpful in future applications such as urban and transportation planning, the sociology of mobility behavior, and spreading dynamics.

In future work, we plan to study the time dimension to propose novel indicators and analysis methods for semantic sequences in a *time-structured approach*. For example, we could describe each activity according to its start and end timestamps. Additionally, we hope to expand our methodology to account multidimensional semantic sequences. Integrating the time dimension and multidimensional semantics will facilitate the treatment of more detailed sequences and enhance our methodology.

## References

- [1] A. Abbott and A. Tsay. Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research*, 29(1):3–33, 2000.
- [2] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.
- [3] L. Alessandretti, P. Sapiezynski, V. Sekara, S. Lehmann, and A. Baronchelli. Evidence for a conserved quantity in human mobility. *Nature Human Behaviour*, 2(7):485–491, 2018.
- [4] H. Alt and M. Godau. Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 05(01n02):75–91, 1995.
- [5] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.
- [6] AUDIAR. Enquête ménages-déplacements en ille-et-vilaine 2018. Technical report, AUDIAR Rennes, 2019.
- [7] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.
- [8] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1 – 74, 2018. Human mobility: Models and applications.
- [9] M. Barthelemy. The statistical physics of cities. *Nature Reviews Physics*, 1(6):406–415, 2019.
- [10] M. Batty. *The New Science of Cities*. The MIT Press, Cambridge, MA, 2013.
- [11] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali. Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518, 2012.
- [12] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.
- [13] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81:591–646, 2009.
- [14] CERTU. Guide méthodologique des enquêtes ménages déplacement. Technical report, CERTU, 2008.
- [15] L. Chen, M. T. Özsü, and V. Oria. Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’05, pages 491–502, New York, NY, USA, 2005. Association for Computing Machinery.
- [16] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. Pastore y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini, and A. Vespignani. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 368(6489):395–400, 2020.
- [17] H. Cramér. *Mathematical methods of statistics*, volume 43. Princeton university press, 1999.
- [18] C. H. Elzinga and M. Studer. Spell sequences, state proximities, and distance metrics. *Sociological Methods & Research*, 44(1):3–47, 2015.
- [19] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
- [20] Eurostat. Harmonised european time use surveys. Technical report, Eurostat., 2019.
- [21] M. C. Ferreira de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):378–394, 2003.
- [22] C. Ferrero, L. Alvares, and V. Bogorn. Multiple aspect trajectory data analysis: Research challenges and opportunities. *GeoInformatica*, 17:56–67, 2016.

- [23] C. A. Ferrero, L. M. Petry, L. O. Alvares, C. L. da Silva, W. Zalewski, and V. Bogorny. Mastermovelets: discovering heterogeneous movelets for multiple aspect trajectory classification. *Data Mining and Knowledge Discovery*, 34(3):652–680, 2020.
- [24] M. Friendly. Mosaic displays for multi-way contingency tables. *Journal of The American Statistical Association*, 89:190–200, 1994.
- [25] A. S. Furtado, D. Kopanaki, L. O. Alvares, and V. Bogorny. Multidimensional similarity measuring for semantic trajectories. *Transactions in GIS*, 20(2):280–298, 2016.
- [26] A. Gabadinho, G. Ritschard, N. S. Mueller, and M. Studer. Analyzing and visualizing state sequences in r with traminer. *Journal of Statistical Software*, 40(4):1–37, 2011.
- [27] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20(5):695, 2011.
- [28] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’07, pages 330–339, New York, NY, USA, 2007. Association for Computing Machinery.
- [29] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- [30] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.
- [31] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 2018.
- [32] S. J. Haberman. The analysis of residuals in cross-classified tables. *Biometrics*, 29(1):205–220, 1973.
- [33] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145, 2001.
- [34] B. Halpin. Optimal matching analysis and life-course data: The importance of duration. *Sociological Methods & Research*, 38(3):365–388, 2010.
- [35] M. Hollister. Is optimal matching suboptimal? *Sociological Methods & Research*, 38(2):235–264, 2009.
- [36] S. Jiang, J. Ferreira, and M. C. González. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, 25(3):478–510, 2012.
- [37] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. González. The timegeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences*, 113(37):E5370–E5378, 2016.
- [38] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. 2009.
- [39] E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386, 2005.
- [40] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *IEEE Transactions on Information Theory*, 44(3):1319–1327, 1998.
- [41] W. T. Krzanowski and Y. T. Lai. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44(1):23–34, 1988.
- [42] A. L. Lehmann, L. O. Alvares, and V. Bogorny. SMSM: a similarity measure for trajectory stops and moves. *International Journal of Geographical Information Science*, 33(9):1847–1872, 2019.
- [43] L. Lesnard. Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3):389–419, 2010.
- [44] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966.
- [45] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS ’08, New York, NY, USA, 2008. Association for Computing Machinery.
- [46] P. G. Lind, L. R. da Silva, J. S. Andrade, and H. J. Herrmann. Spreading gossip in social networks. *Phys. Rev. E*, 76:036117, 2007.
- [47] M. Lv, L. Chen, and G. Chen. Mining user similarity based on routine activities. *Information Sciences*, 236:17–32, 2013.

- [48] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605, 2008.
- [49] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [50] B. D. McKay and A. Piperno. Practical graph isomorphism, ii. *Journal of Symbolic Computation*, 60:94 – 112, 2014.
- [51] A. Menin, S. Chardonnell, P.-A. Davoine, and L. Nedel. eSTIME: Towards an All-in-One Geovisualization Environment for Daily Mobility Analysis. In *32nd Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 39–46, 2019.
- [52] C. Moreau, T. Devogele, and E. Laurent. Contextual edit distance for semantic trajectories. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 635–637, 2020.
- [53] C. Moreau, V. Peralta, P. Marcel, A. Chanson, and T. Devogele. Learning analysis patterns using a contextual edit distance. In *DOLAP 2020, EDBT/ICDT*, volume 2572, pages 46–55, 2020.
- [54] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in NIPS*, pages 849–856, 2002.
- [55] G. Pan, G. Qi, Z. Wangsheng, S. Li, Z. Wu, and L. T. Yang. Trace analysis and mining for smart cities: issues, methods, and applications. *IEEE Communications Magazine*, 51(6):120–126, 2013.
- [56] L. Pappalardo and F. Simini. Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery*, 32(3):787–829, 2018.
- [57] L. Pappalardo, F. Simini, G. Barlacchi, and R. Pellungrini. scikit-mobility: A python library for the analysis, generation and risk assessment of mobility data. *arXiv preprint arXiv:1907.07062*, 2019.
- [58] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan. Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4), 2013.
- [59] H.-S. Park and C.-H. Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.
- [60] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Rev. Mod. Phys.*, 87:925–979, 2015.
- [61] E. Pebesma. Cran task view: Handling and analyzing spatio-temporal data.
- [62] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- [63] G. Rohwer and P. Ulrich. Tda user's manual. Technical report, Universität Bochum, 2005.
- [64] L. Rokach and O. Maimon. *Clustering Methods*, pages 321–352. 2005.
- [65] P. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
- [66] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.
- [67] C. Song, T. Koren, P. Wang, and A.-L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.
- [68] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- [69] A. Struyf, M. Hubert, and P. J. Rousseeuw. Clustering in an object-oriented environment. *Journal of Statistical Software*, 1(4):652–680, 1997.
- [70] M. Studer and G. Ritschard. What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 179:481–511, 02 2016.
- [71] D. d. C. Teixeira, A. C. Viana, M. S. Alvim, and J. M. Almeida. Deciphering predictability limits in human mobility. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 52–61, 2019.

- [72] M. Vlachos, G. Kollios, and D. Gunopulos. Discovering similar multidimensional trajectories. In *Proceedings 18th International Conference on Data Engineering*, pages 673–684, 2002.
- [73] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, 1974.
- [74] C. Wan, Y. Zhu, J. Yu, and Y. Shen. Smopat: Mining semantic mobility patterns from trajectories of private vehicles. *Information Sciences*, 429:12 – 25, 2018.
- [75] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.
- [76] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, ACL '94*, pages 133–138, USA, 1994. Association for Computational Linguistics.
- [77] C. Zhang, J. Han, L. Shou, J. Lu, and T. Porta. Splitter: Mining finegrained sequential patterns in semantic trajectories. *Proceedings of the VLDB Endowment*, 7:769–780, 05 2014.
- [78] G. Zhu and C. A. Iglesias. Computing semantic similarity of concepts in knowledge graphs. *IEEE Trans. on Knowledge and Data Engineering*, 29(1):72–85, 2016.