Summarizing Trajectories Using Semantically Enriched Geographical Context

Chiara Pugliese
ISTI-CNR, University of Pisa, Italy
chiara.pugliese@isti.cnr.it

Francesco Lettich
ISTI-CNR, Pisa, Italy
francesco.lettich@isti.cnr.it

Fabio Pinelli
IMT Lucca, Italy
fabio.pinelli@imtlucca.it

Chiara Renso
ISTI-CNR, Pisa, Italy
chiara.renso@isti.cnr.it

ABSTRACT

The proliferation of tracking sensors in today's devices has led to the generation of high-frequency, high-volume streams of mobility data capturing the movements of various objects. These movement data can be enriched with semantic contextual information, such as activities, events, user preferences, and more, generating semantically enriched trajectories. Creating and managing these types of trajectories presents challenges due to the massive data volume and the heterogeneous, complex semantic dimensions. To address these issues, we introduce a novel approach, MAT-Sum, which uses a location-centric enrichment perspective to summarize massive volumes of mobility data while preserving essential semantic information. Our approach enriches geographical areas with semantic aspects to provide the underlying context for trajectories, enabling effective data reduction through trajectory summarization. In the experimental evaluation, we show that MAT-Sum effectively minimizes trajectory volume while retaining a good level of semantic quality, thus presenting a viable solution to the relevant issue of managing massive mobility data.

CCS CONCEPTS

• **Information systems** → **Location based services**; **Geographic information systems**.

KEYWORDS

Semantic trajectory, multiple aspect trajectory, summarized semantic trajectory, semantic enrichment

ACM Reference Format:
Chiara Pugliese, Francesco Lettich, Fabio Pinelli, and Chiara Renso. 2023. Summarizing Trajectories Using Semantically Enriched Geographical Context . In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3589132.3625587

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0168-9/23/11.
https://doi.org/10.1145/3589132.3625587

1 INTRODUCTION

In today's world, an abundance of mobility data is generated by devices equipped with tracking sensors. Moreover, there is an important line of research [2, 7, 16, 17, 21, 23] that enriches such data with diverse semantic information (i.e., *aspects*), thereby resulting in semantic or multiple aspect trajectories. The existing approaches present several challenges. Firstly, the data generated can be massive, characterized by high sampling rates, while the semantic aspects can be heterogeneous and have a large number of associated attributes. Consequently, the resulting complexity requires significant storage resources and computational capabilities to ensure effective analysis and meaningful interpretation. To overcome these challenges, researchers have focused on developing methods to reduce the volume of mobility data through summarization or simplification of trajectories [1, 8].

This paper aims to summarize the vast amount of mobility data while preserving the semantic information contained within. We propose a location-centric enrichment perspective, which complements the trajectory-centric approach of the state-of-the-art, to achieve the summarization objective. This perspective enables spatio-temporal and semantic information aggregation, resulting in a more concise representation of mobility data. Moreover, our approach plays a crucial role in enriching mobility data by emphasizing geographical contextual knowledge. Indeed, the combination of these two perspectives provides a more comprehensive understanding of trajectories. For example, by incorporating geographical context into check-in information, we can determine that an individual visits a restaurant (trajectory-centric) situated in an area with numerous tourist attractions and a pedestrian-friendly environment (location-centric). This integration enhances our understanding of the mobility patterns and factors influencing them.

To address the aforementioned objectives, we aim to answer the following research questions:

RQ1 How can we devise a trajectory summarization method that meaningfully leverages the semantic context of the underlying geographical area?
RQ2 To what extent is the summarization method effective in summarizing semantic trajectories?

We begin our investigation from the first research question and address the underlying problem by introducing the MAT-Sum method. MAT-Sum's key idea is to consider the geographical area where trajectories move and use it both for their semantic enrichment and summarization. To this end, MAT-Sum first partitions the

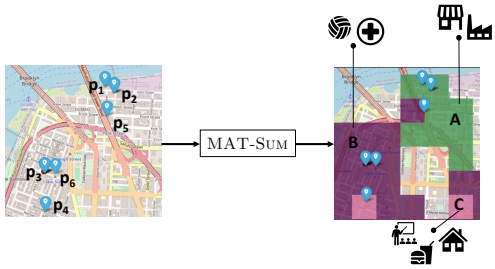


Figure 1: Overview of the MAT-Sum method: from a trajectory and a geographical area to a set of semantic locations.

geographical area into cells, and then enriches each cell with semantic aspects. Subsequently, MAT-Sum leverages the enriched cells to obtain the summarized semantic trajectories from the initial ones. More specifically, each trajectory – which might have been already semantically enriched with a trajectory-centric approach or not – is discretized into the set of enriched cells it traverses. During the discretization, the cells serve a dual purpose: they not only enrich the trajectory with their semantic contexts in a location-centric way but are also used to condense the information concerning its movements and any semantic information it might possess. Figure 1 provides a schematic representation of the input and output of MAT-Sum. On the left, a trajectory is depicted as a sequence of points, p_1 through p_6, traversing some geographical area. Starting from these points, MAT-Sum derives a set of semantically enriched geographical areas, each endowed with a weight indicating the temporal importance of the location for the trajectory. We call the final result a *summarized semantic trajectory*.

For what concerns the second research question, in the experimental evaluation we assess MAT-Sum's effectiveness by considering two orthogonal and contrasting aspects: minimizing the information contained in the summarized semantic trajectories while simultaneously preserving an adequate level of semantic quality. Consequently, we evaluate our approach with two distinct datasets over different scenarios and compare its ability to obtain high-quality summarized semantic trajectories w.r.t. two baseline methods, i.e., RLE and Seqscan-D [6]. Overall, we show that MAT-Sum successfully accomplishes its goals.

The rest of the paper is structured as follows. In Section 2, we present the related works and highlight the novelty behind our approach. In Section 3, we provide the fundamental concepts used throughout the paper and the problem definition. Section 4 details the MAT-Sum approach. In Section 5, we provide the experimental evaluation. Finally, Section 6 draws the conclusions.

2 RELATED WORK

In this paper, we consider the problem of summarizing semantic trajectories. The related works therefore fall into two distinct research fields: *semantic enrichment of trajectories* and *trajectory summarization*. The first field focuses on the problem of enriching movement data with additional contextual information. The second one encompasses approaches that are centred on the simplification and abstraction of movement data, where the aim is to retain essential information about the original trajectories, while significantly reducing data volume.

The concept of semantic trajectories has been introduced in the seminal work by Spaccapietra et al. [23], where trajectories are partitioned into *stop segments*, i.e., sub-trajectories where the moving objects remain stationary, and *move segments*, i.e., sub-trajectories where the objects alter their positions. Several subsequent studies sought to provide more formalization and more types of enriching data, thus going beyond the simple stop and move paradigm. For instance, [2] introduces additional semantic dimensions, [7] and [17] present domain ontologies for semantically enriched trajectories, and [21] introduces a method to create enriched trajectories from Linked Open Data. A more recent paper proposes MASTER [16], a conceptual model for enriching trajectories with multiple *aspects* – that is, semantic dimensions that can be heterogeneous and have complex representations. Lastly, a few recent works [12, 20] proposed a general methodology and a system that lets users incorporate the aspects and the data sources they need in their own semantic enrichment processes. All the above approaches are based on the concept of trajectory segmentation, wherein segments (sub-trajectories) are first determined based on certain criteria (e.g., stops and moves [23]) and then enriched with some aspects. We call such enrichment perspective *trajectory-centric*. In this work, we adopt a different and complementary enrichment perspective, which places less emphasis on the trajectories themselves and, instead, is centred on the contextual semantic information provided by the geographical areas through which moving objects navigate. We call this perspective *location-centric*. It is crucial to understand that both perspectives can work together synergistically. In the location-centric perspective, we are not only enriching trajectories with the underlying geographical context but we also use the same context to summarize them. Therefore, we must also consider related works in the field of *data summarization*.

Generally speaking, data summarization is a mining process that transforms data into a concise and informative representation, possibly abstracting content from the original data [3]. As such, data summarization goes beyond simple compression (see, for instance, the techniques surveyed in [11]). While summarization might lead to approximate results when compared to the original data, it offers substantial savings in time and space, making it a practical approach for handling large datasets while retaining pertinent information. When the data in question is trajectory data, we enter the sub-field of *trajectory summarization*. Trajectory summarization is the task of finding a compact representation of the spatio-temporal movement of objects while preserving the relevant information [1]. This is an established research field with numerous proposed methods for reducing trajectory points, such as trajectory compression, simplification, or segmentation [8]. Trajectory simplification (or trajectory cleaning) refers to the task of reducing the number of samples when the sampling rate is high while preserving the trajectory's spatio-temporal characteristics [8]. Alternatively, for certain types of movement data, good compression rates can be achieved by using the underlying road networks with little or no reduction in quality [26]. Other works simplify trajectories based on some type of semantics [4, 8, 13, 24]. These methods are related to our work since they aim at reducing trajectory size by exploiting the underlying semantic information. However, in contrast to our approach they do not aim to preserve semantics, but instead use it to simplify trajectories.

To the best of our knowledge, the existing literature on the summarization of semantic trajectories includes works such as [5, 6, 15]. Damiani et al. [5, 6] propose a trajectory summarization method that extracts locations of interest from symbolic trajectories (these can be seen as simple types of semantic trajectories). The method is tailored to telecommunications data, combining location attractiveness and frequency to classify the visited places. Furthermore, it takes into account the diversity of the locations of interest. The authors evaluate it using a large telecommunications dataset, demonstrating a significant reduction in data complexity while still delivering high-quality information on mobility behaviors.

In a recent paper, Machado et al. [15] introduce the problem of computing representatives for groups of multiple aspect trajectories, which can be seen as a form of summarization. The authors assess the effectiveness of their approach through an evaluation of volume reduction and accuracy. Similarly to our approach, they propose a grid-based trajectory summarization.

Overall, while our approach and those from [15] and [6] all involve trajectory summarization, there are important distinctions to be made. Unlike [6], MAT-Sum ensures that no traversed location is discarded, and it also leverages geographical semantic context. Moreover, MAT-Sum operates at the individual trajectory level, while [15] focuses on generating a representative trajectory for predefined groups of trajectories (i.e., clusters), thereby creating a single representative structure for each group.

3 PRELIMINARIES AND PROBLEM DEFINITION

In this section, we introduce the basic concepts used throughout the rest of the paper and then define the problem we aim to solve. Let us first introduce the notion of *aspect* [16], which is needed for formalizing any kind of semantically enriched entity (i.e., trajectories and geographical areas).

Definition 3.1 (Aspect). An aspect is a set $A = \{a_1, a_2, ..., a_l\}$ of l characterizing attributes that semantically represent A. We define the *instance* of an aspect A a specific instantiation of its attributes.

An aspect is essentially any sort of information that can be used to annotate an entity. For instance, we may define aspects such as POI, weather, and transportation means, each represented semantically by distinct attributes (e.g., POIs are described by type, rating, and price tier attributes). When aspects are used to semantically enrich trajectory segments (i.e., sub-trajectories), we have the concept of *multiple aspect trajectory*. Its formal definition has been first introduced in [16]. In the context of our work, we use the slightly simplified version of the definition from [19].

Definition 3.2 (Multiple aspect trajectory). A multiple aspect trajectory is a sequence of points $T = \langle p_1, p_2, \ldots, p_n \rangle$, with $p_i = (x_i, y_i, t_i, ASP_i)$ being the i-th point of the trajectory at the time-stamped location (x_i, y_i, t_i) enriched with instances of the aspects in $ASP_i = \{A_1, A_2, \ldots, A_r\}$.

Observe that Definition 3.2 can encompass several types of trajectories, ranging from the raw trajectories – in this case ASP_i is always an empty set – to various types of semantically enriched trajectories with different degrees of complexity, e.g., [2, 6, 9, 16, 23]. For simplicity, henceforth we will use the general term trajectories to indicate all types of trajectories (e.g., raw or semantically enriched) and the term semantic trajectory to refer to any trajectory that has been semantically enriched.

3.1 Problem definition

From a dataset of trajectories D, we intend to derive a dataset of summarized semantic trajectories D_{SumSem}. Central to this problem is the concept of *geographical context*, which we define as the geographical area the trajectories traverse, augmented with selected semantic aspects. This geographical context serves the dual purpose of enriching and summarizing the trajectories. Regardless of whether the trajectories in dataset D are already trajectory-centric enriched or not, it is essential for the summarization process to preserve both the original and the geographic context-derived semantics while reducing the data volume.

When generating D_{SumSem}, we are therefore faced with two potentially conflicting objectives that must be balanced: on one hand, we want to enrich the initial trajectories with semantics provided by the underlying geographical context. On the other hand, we aim to reduce the information within the summarized trajectories, still leveraging the underlying geographical context.

First, let us define the function in charge of enriching a geographical area of interest with selected aspects.

Definition 3.3 (Map enrichment). Let $\mathcal{G}$ be the space of all geographical areas, $\mathcal{ASP}$ be the space of all aspects that can be used to enrich geographical areas, and $\mathcal{G}_{Sem}$ be the space of all possible enriched geographical areas, i.e., geographical contexts. Then, $MapEnrich : \mathcal{G} \times \mathcal{ASP}^n \rightarrow \mathcal{G}_{Sem}$ is a function that enriches the areas in $\mathcal{G}$ with n semantic aspects, thus yielding a geographical context.

Next, we provide a generic definition of the function enriching the initial trajectories with the underlying geographical context.

Definition 3.4 (Trajectory enrichment with geographical context). Let $\mathcal{T}$ be the space of the trajectories to be summarized, and $\mathcal{G}_{Sem}$ be the space of all possible enriched geographical areas. Finally, let $\mathcal{T}_{GeoSem}$ be the space of the trajectories enriched with the underlying geographical context. Then, $Enrich : \mathcal{T} \times \mathcal{G}_{Sem} \rightarrow \mathcal{T}_{GeoSem}$ is a function that enriches the trajectories in $\mathcal{T}$ with instances of the aspects enriching $\mathcal{G}_{Sem}$, thus equipping the trajectories with the underlying geographical context.

It is important to note that *Enrich* does not destroy any information within the initial trajectories. Next, we provide the generic definition of the summarization function.

Definition 3.5 (Trajectory summarization). Let $\mathcal{T}_{GeoSem}$ be the space of all possible trajectories enriched with the underlying geographical context and $\mathcal{T}_{SumSem}$ be the space of summarized semantic trajectories. Then, $Summarize : \mathcal{T}_{GeoSem} \rightarrow \mathcal{T}_{SumSem}$ is a function that summarizes the enriched trajectories.

We can now formalize the two aforementioned contrasting objectives behind the problem of summarizing semantic trajectories. First, let us define in generic terms the function in charge of measuring how much a summarized semantic trajectory condenses the information of the trajectory from which it is derived.

Definition 3.6 (Summarization rate). Let $\mathcal{T}_{GeoSem}$ be the space of all possible trajectories enriched with the underlying geographical context, and $\mathcal{T}_{SumSem}$ be the space of all possible summarized semantic trajectories. Then, $SumRate : \mathcal{T}_{GeoSem} \times \mathcal{T}_{SumSem} \rightarrow [0, 1]$ is a function that assesses the extent to which a summarized semantic trajectory reduces the information in the original enriched trajectory. A value close to 1 indicates a more substantial reduction.

Let us now formalize the other objective, which concerns the quality of the semantic information retained in the summarized trajectories.

Definition 3.7 (Semantic quality). Let $\mathcal{T}_{GeoSem}$ be the space of all possible trajectories enriched with the underlying geographical context, and $\mathcal{T}_{SumSem}$ be the space of all possible summarized semantic trajectories. Then, $SemQual : \mathcal{T}_{GeoSem} \times \mathcal{T}_{SumSem} \rightarrow [0, 1]$ is a function that evaluates how well the semantic similarity has been preserved in a summarized trajectory, where values close to 1 indicate better results.

We now formalize the problem we intend to solve.

Definition 3.8 (Problem Definition). Let D be a dataset of trajectories, and G be the geographical area in which the trajectories move. Moreover, let $G_{Sem} = MapEnrich(G, \{A_1, \ldots, A_n\})$ be the enriched geographical area, and $D_{GeoSem} = \{Enrich(T, G_{Sem}) \mid T \in D\}$ be the dataset of trajectories enriched with the underlying geographical context. Then, we seek to find a *Summarize* function that concurrently maximizes the summarization rate and the semantic similarity over the entire dataset. Assume that $\hat{T} = Summarize(T)$ for any $T \in D_{GeoSem}$; then, we want to solve:

$$\underset{Summarize}{\arg\max} \sum_{T \in D_{GeoSem}} SumRate(T, \hat{T}) + SemSim(T, \hat{T}). \quad (1)$$

In Section 4, we introduce MAT-Sum, the method we propose for implementing the *MapEnrich*, *Enrich*, and *Summarize* functions, while in Section 5, we instantiate the *SumRate* and *SemQual* measures and use them to assess MAT-Sum's effectiveness.

4 THE MAT-SUM APPROACH

This section introduces MAT-Sum, our trajectory summarization proposal. The method consists of two phases, depicted in Figure 2. The initial phase, *map semantic enrichment* (Section 4.1), enriches the geographical area from a location-centric view, creating semantic locations. This process involves tessellating the area, assigning a semantic context to each tile, and merging those tiles that share identical contexts. This yields a set of semantic locations. The subsequent phase, *semantic trajectory summarization* (Section 4.2), uses the semantic locations to summarize trajectories. Here, each trajectory is transformed into a weighted set of traversed semantic locations, each enriched with temporal weights and original semantic attributes. Overall, the two phases address the objectives outlined in the problem definition (Definition 3.8 in Section 3.1).

4.1 Map semantic enrichment

Starting from a geographical area of interest (i.e., the one in which trajectories move), the goal of this phase is to obtain a set of semantic locations. This is achieved through three steps: **map tessellation**, **tile semantic enrichment**, and **enriched tiles union**, as illustrated in the upper section of Figure 2.

Map tessellation. The input of this step is the geographic area covered by the trajectories and, through the use of some tessellation method, produces a set of tiles covering the area. Formally, let G be the geographical area defined by a rectangle. Then, we recall that a spatial tessellation superimposed over G is a set of tiles $L = \{l_1, \ldots, l_n\}$ that are collectively exhaustive and mutually exclusive, except for the boundaries [18]. In the literature, several types of tessellations exist, ranging from uniform, hexagonal, Voronoi, and city-block, as well as adaptive types of tessellations based on some notion of object density such as adaptive squares and quad-trees [10, 18, 22]. The set of tiles is the next step's input.

Tile semantic enrichment. The goal of this step is to add contextual, geographical information to the tiles, thereby converting them into semantically enriched tiles. Accordingly, we address this problem as follows: given some tessellation (e.g., uniform grid) and a set of semantic aspects, we obtain a set of semantically enriched tiles by adding to each tile the semantic information pertinent to the chosen aspects and the geographical area it encompasses.

Definition 4.1 (Enriched tiles). Let $L = \{l_1, \ldots, l_n\}$ be the set of tiles previously computed, and $ASP = \{A_1, \ldots, A_m\}$ a set of semantic aspects, each having their own set of aspect instances (as per Definition 3.1). Here, we require that said aspects have an attribute that provides the geographical location of their instances. We then carry out a *spatial join* between the tiles in L and the instances belonging to the aspects in ASP. This operation yields $L' = \{(l_1, S_1), \ldots, (l_n, S_n)\}$, where each tile l_i is enriched with S_i, i.e., the set of aspect instances geolocated in l_i and that provide the tile's geographical semantic context.

Enriched tiles union. After obtaining L', the semantically enriched tiles can be distinguished by the associated semantic context. As MAT-Sum prioritizes semantic information, it is important to note that areas within a city may exhibit the same semantic characteristics. To account for this, we merge semantically enriched tiles that share the same semantic context, i.e., those that have identical instances of semantic aspects associated with them, thus leading to *semantic locations*. We highlight that during the merging process, the geographic coordinates associated with the aspect instances are not taken into account. More formally:

Definition 4.2 (Semantic location). Given a set of enriched tiles $L' = \{(l_1, S_1), \ldots, (l_n, S_n)\}$, we define a semantic location sl as the geographical union of all enriched tiles in L' that share the same set of aspect instances $\bar{S}$. Formally:

$$sl = (\bigcup_{(l,S) \in L'} l\ ,\ \bar{S}\ ), \text{ such that } S = \bar{S},$$

Note that a semantic location can be composed of non-adjacent tiles, thus resulting in a multi-polygon region. We denote the set of semantic locations by $L'' = \{sl_1, \ldots, sl_m\}$, where $m \leq n$.

Ultimately, the geographical area of interest is partitioned into a set of semantic locations, achieved by merging semantically enriched cells according to their respective semantic contexts. Overall, this phase implements the *MapEnrich* function introduced in Definition 3.3 and yields a geographical context that can be used for the enrichment and summarization of trajectories in the next phase.

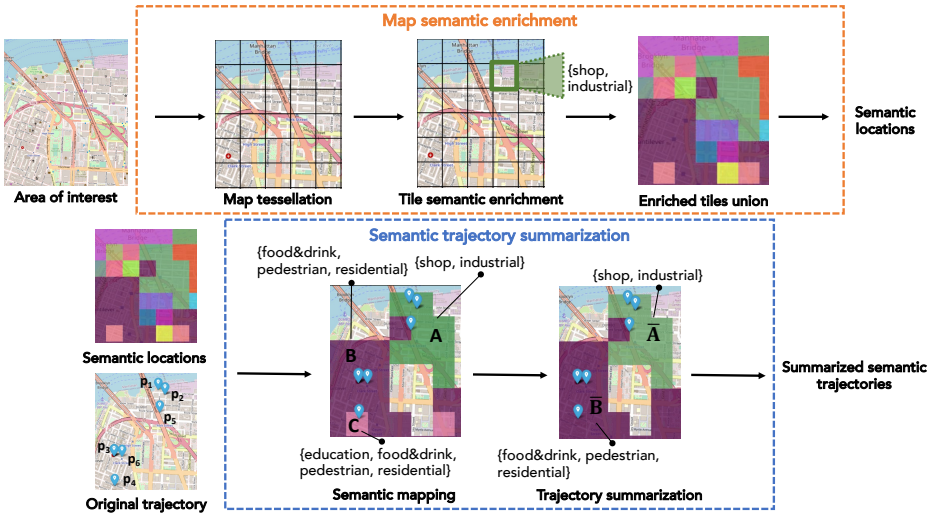


Figure 2: The MAT-Sum approach organized into two main phases: the map semantic enrichment depicted in the orange box at the top and the trajectory summarization illustrated in the blue box at the bottom.

4.2 Semantic trajectory summarization

In this phase, given a trajectory and the set of semantic locations obtained in the previous phase, the goal is to enrich the trajectory with the underlying geographical context and use the same context to summarize it, i.e., reduce the number of semantic locations it traverses based on their similarities. This phase is illustrated in the lower section of Figure 2 and consists of two steps.

The first step implements the *Enrich* function (see Definition 3.4) and involves transforming the trajectory into a temporally weighted sequence of semantic locations it traverses. It is important to note that the trajectory could either be raw or already enriched from a trajectory-centric perspective. This conversion ensures that the trajectory is represented in terms of meaningful semantic locations. We call this step **semantic mapping**.

In the second step, which implements the *Summarize* function (see Definition 3.5), the semantic locations traversed by the trajectory are aggregated based on their similarities. This aggregation process is integral to the summarization: by grouping similar semantic locations together, the trajectory can be represented with fewer locations while still retaining essential semantic information. We call this step **trajectory summarization**.

Collectively, the two steps enable the semantic trajectory summarization phase to output summarized trajectories that maintain semantic quality while significantly reducing data volume.

Semantic mapping. This first step is depicted in the bottom box of Figure 2 and aims to associate the original trajectory with the corresponding semantic locations and temporal weights.

Definition 4.3 (Weighted sequence of semantic locations). Let $T = \langle p_1, \ldots, p_n \rangle$ be a trajectory (see Definition 3.2). Recall that in the case of a raw trajectory, the *ASP* term in each $p \in T$ is empty. Then, by means of a spatial join, we associate each point $p \in T$ with the semantic locations in L'' (see Definition 4.2). This converts T into a **weighted sequence of semantic locations**. We denote such sequence by $\tilde{T} =< (p_1, sl_1, w_1), \ldots, (p_n, sl_n, w_n) >$, where w_i is the *temporal weight* obtained considering the time the moving object spent to traverse the semantic location sl_i associated with the original point p_i. We emphasize that in $\tilde{T}$ the original trajectory-centric semantics from T are preserved. This implements the *Enrich* function in Definition 3.4 of the problem definition.

It is important to note that the temporal weight w_i represents temporal duration and is measured in units of time (such as seconds or minutes). This is particularly relevant in the context of dense trajectories (e.g., GPS trajectories), where it is feasible to calculate the time interval between two consecutive points, while we set it equal to 1 in the case of sparse trajectories (i.e., check-ins).

Trajectory summarization. Once the trajectory is transformed into a weighted sequence of semantic locations, it is possible to proceed with its summarization (see Definition 3.5). As highlighted in the previous section, assigning a set of semantic aspect instances to each location enables identifying the characteristics of the areas traversed by the trajectories. Moreover, we aim to facilitate the summarization by leveraging some characteristics of human mobility behaviour. For instance, individuals often tend to visit similar places, and human movement patterns are generally predictable [14, 25]. With this in mind, we can merge two or more locations into a single one if they share a significant number of aspect instances. This number is determined based on a predefined threshold. Formally, given a trajectory transformed into a weighted sequence of semantic location, $\tilde{T}$, and any pair of semantic locations $sl_j, sl_i \in \tilde{T}$, with $sl_j \neq sl_i$, we state that sl_j is *similar* to sl_i, and we denote it by $sl_j \sim sl_i$ if the following holds:

$$sl_j \sim sl_i \text{ if } sim(S_j, S_i) \geq \tau, \tag{2}$$

where S_j and S_i are respectively the semantic contexts associated with sl_j and sl_i, $sim(\cdot, \cdot)$ is a similarity function applicable to pairs of such sets and that outputs a value in $[0, 1]$ (e.g., cosine similarity, Jaccard similarity), and τ is some similarity threshold value.

The similarity measure, once computed for each pair of distinct semantic locations in $\tilde{T}$, leads to the notion of *similarity class*. Let sl_i be a semantic location belonging to $\tilde{T}$. Then, we denote by $[sl_i]$ the similarity class of sl_i, which we define to be the set of semantic locations similar to sl_i, and is given by: $[sl_i] = \{sl_j \in \tilde{T} \mid sl_j \sim sl_i\}$. For each similarity class, we can then define its *representative semantic location* $\overline{sl}_i$:

$$\overline{sl}_i = (\bigcup_{sl_j \in [sl_i]} l_j, \bigcap_{sl_j \in [sl_i]} S_j). \quad (3)$$

In other words, a representative semantic location is given by the geographical union of the semantic locations in the similarity class and the set intersection of their semantic contexts. Observe that the definition in Equation 3 leads to a more condensed representation of the semantic contexts of similar locations by capturing only the shared aspects. Finally, we denote by $SL = \{\overline{sl_1}, \ldots, \overline{sl_m}\}$ the set of representative semantic locations, where $m \leq n$.

We would like to emphasize that an alternative approach could be to directly integrate a similarity function into the definition of semantic location (see Definition 4.2), which would merge semantic locations with similar, albeit not necessarily identical, contexts. However, in this work, we have opted to carry out this operation at the trajectory level. This choice is motivated by the desire to take into account the impact of the similarity threshold τ on the level of summarization. By doing so, we can specifically focus on trajectory summarization while still retaining the original semantic information from the semantic locations. Finally, we proceed with the implementation of the *Summarize* function from Definition 3.5.

Definition 4.4 (Summarized semantic trajectory). Given a weighted sequence of semantic locations $\tilde{T}$ and the set of representative semantic locations SL previously defined, we aggregate each point in $\tilde{T}$ that falls in the same representative semantic location $\overline{sl}_i \in SL$. The resulting **summarized semantic trajectory** $\hat{T}$ is given by:

$$\hat{T} = \{(\overline{sl}_i, \bigcup_{sl_j \in [sl_i]} ASP_j, \sum_{sl_j \in [sl_i]} w_j) \mid \overline{sl}_i \in SL \wedge (sl_j, ASP_j, w_j) \in \tilde{T}\},$$

where $\overline{sl}_i$ is one of the representative semantic locations making up $\hat{T}$, $\bigcup_{sl_j \in [sl_i]} ASP_j$ is the union of the aspect instances associated with the points of the original trajectory that fall within the representative semantic location $\overline{sl}_i$, and $\sum_{sl_j \in [sl_i]} w_j$ is the sum of temporal weights representing the total time spent in all semantic locations belonging to the same similarity class.

Intuitively, a summarized semantic trajectory $\hat{T}$ is comprised of a set of semantic locations wherein the context is preserved yet distilled as previously explained. In $\hat{T}$, each semantic location is enriched with the union of aspect instances (if any) from the enriched trajectory $\tilde{T}$. This process can be seen as a distillation of the semantic aspects, disregarding individual occurrences of aspect instances while preserving the underlying information. Additionally, each semantic location is assigned a weight that corresponds to the cumulative sum of weights representing the time spent in all analogous areas.

Example. Consider a simple raw trajectory T, comprising of six timestamped geographical check-ins (e.g., recorded in New York), i.e., $T = \langle p_1, p_2, p_3, p_4, p_5, p_6 \rangle$. We recall that the set of semantic aspects associated with each p_i is empty in the case of a raw trajectory. As shown in Figure 2, we identify New York as the area of interest and, consequently, determine the bounding box. We then tessellate the box with a uniform grid. Once we obtain the set of semantic locations (as described in Definition 4.2), we perform a spatial join (see the semantic mapping step in Figure 2) between the points in T and the semantic locations, yielding the weighted sequence of semantic locations:

$$\tilde{T} = \langle (p_1, A, 1), (p_2, A, 1)(p_3, B, 1), (p_4, C, 1)(p_5, B, 1)(p_6, B, 1) \rangle,$$

where p_1 and p_2 fall within the same semantic location A; p_3, p_5, and p_6 fall within B; p_4 falls in C.

Let us now consider the semantic contexts associated with the semantic locations A, B, and C, and defined as follows: $S^{(A)}$ = {shop, industrial}, $S^{(B)}$ = {food and drink, pedestrian, residential}, $S^{(C)}$ = {education, food and drink, pedestrian, residential}. We can aggregate the semantic locations in $\tilde{T}$ according to Definition 4.4, weighting them with the number of semantic location occurrences. Since the semantic contexts associated with the semantic location B and C are very similar, and supposing that $sim(B, C) \geq \tau$ (e.g., $\tau = 0.9$), the new summarized semantic location is $\hat{T} = \{(\overline{A}, 2)(\overline{B}, 4)\}$, where $\overline{B}$ is the representative semantic location of the similarity class of $[B]$, and $\overline{A}$ of $[A]$, obtained as described in Equation 3.

We highlight that the semantic context associated with $\overline{B}$ is {*food and drink, pedestrian, residential*}, i.e., the intersection between $S^{(B)}$ and $S^{(C)}$. It is important to note that even if $\tau = 1$, resulting in similarity being defined as equality, we obtain the following summarized semantic trajectory: $\hat{T} = \{(\overline{A}, 2), (\overline{B}, 3), (\overline{C}, 1)\}$.

To sum up, in this section we have presented the MAT-Sum approach to semantically enrich trajectories with geographical contextual information and, at the same time, transform them into a summarized format that can retain the semantics while reducing the data volume. Therefore, MAT-Sum answers the research question RQ1 from Section 1 with a method that properly combines geography and semantics to propose a summarized semantic version of trajectories. In the next section, we describe the evaluation process to answer the Research Question RQ2.

5 EXPERIMENTAL EVALUATION

In the experimental evaluation, we address the research question RQ2 from Section 1 to assess the effectiveness of MAT-Sum in solving the problem introduced in Section 3.1. The details of the experimental setup are introduced in Section 5.1 where we discuss the dataset used, the baselines, and the evaluation measures. In Section 5.2, we present three experimental studies aimed at assessing the effectiveness of MAT-Sum under various combinations of input parameters. The first study concentrates on the summarization rate achieved by our method, the second study focuses on the semantic quality that MAT-Sum attains, while the third study compares our method against the two baselines.

5.1 Experimental setup

In this section, we outline the experimental setup used to evaluate MAT-Sum, namely, the datasets used, the baselines against which we compare the results of our method, and the measures used to assess the summarization rate and semantic quality.

Datasets. We evaluate our method utilizing two distinct datasets of semantically enriched trajectories, and an additional dataset for contextual geographical enrichment. The Foursquare NYC dataset [27] comprises 227,428 check-ins from 1,083 distinct users, collected between April 2012 and February 2013. Each check-in consists of a geographically pinpointed timestamp, enriched with venue information such as category, rating, and price, as well as weather conditions. We adopt the approach suggested by Petry et al. [19] and we divide each user's check-ins into weekly segments. The Geolife dataset [28] contains 17,621 GPS trajectories of 178 users, recorded between April 2007 and October 2011. Each entry in the dataset is a tuple consisting of timestamped geographical coordinates, and in some cases, it is further enriched with transportation mode information. The dataset covers trajectories from 30 cities in China, as well as some locations in the USA and Europe. However, for this study, we focus solely on trajectories within the geographical boundaries of Beijing, China.

Lastly, to semantically enrich tiles (as outlined in Section 4.1) we utilize selected semantic aspects obtained from OpenStreetMap[1] (OSM). Specifically, we download all *points of interest* within the area of interest, as well as *land use* and *public transportation* data, selecting the *category* attribute. Since data retrieved from OSM can be incomplete or inconsistent, due to non-standardized tags and volunteered geographical information, we adopt a fixed list of officially recommended data categories[2]. Consequently, we map all the categories within this list, ensuring that semantic information extracted from OSM remains consistent and unambiguous. For the implementation details on how we use OSM data to enrich tiles, we refer the reader to the MAT-Sum's Github repository[3].

Baselines. To evaluate the effectiveness of MAT-Sum, we conducted a comparative analysis with two baseline methods already used in [6]: Run-Length Encoding (RLE) and Seqscan-D. RLE is a compression technique for sequences. It operates by encoding consecutive runs of identical values in a sequence into pairs (l, w), where l represents the value and w denotes the frequency of that value in the sequence. In the case of a sequence of locations, RLE can be adopted to identify salient locations. For example, given the sequence of locations in the context of trajectory data *ABBBAAACC*, it would be encoded as $(A, 1), (B, 3), (A, 3), (C, 2)$.

Seqscan-D is a trajectory summarization method that utilizes a density-based trajectory segmentation approach tailored specifically for telecommunications data. Its primary goal is to identify dominant locations that become representative locations for specific time periods. Seqscan-D first assigns weights to consecutive identical locations at time t_i and t_{i+1} based on their temporal distance, expressed as $|t_{i+1} - t_i|$. Then, it identifies the dominant locations by leveraging the concept of well-formed subsequences. A location is dominant when the starting and ending locations are the same, the length of this subsequence exceeds a predefined threshold N, and the assigned weight is greater than or equal to a specified value of δ. In the example reported in [6], we have a trajectory homogeneously spaced in time of 2 time units $T = (a, t_1)(a, t_2)(c, t_3)(a, t_4)(c, t_5)(b, t_6)(b, t_7)(a, t_8)(b, t_9)(b, t_{10})$, if we set $N = 3$ and $\delta = 2$, the resulting summarized trajectory will be $\hat{T} = ([t_1, t_4], a)([t_6, t_{10}], b)$. If, on the other hand, we set $\delta = 4$, the summarized trajectory will be $\hat{T} = ([t_6, t_{10}], b)$. It is worth noting that the trajectory summarization approach used by Seqscan-D differs from the approach employed in MAT-Sum. Our method is more general and does not focus on summarizing telco trajectory data by identifying dominant locations.

Evaluation measures and semantic similarity. To assess the effectiveness of MAT-Sum, we need to determine the extent to which our method can simultaneously maximize the *summarization rate* and the *semantic quality* measures, as outlined in Definition 3.8. The two measures will be used to compare weighted sequences of semantic locations (Definition 4.3) with their summarized counterparts (Definition 4.4).

We instantiate the summarization rate measure as follows. Let $\tilde{T}$ be a weighted sequence of semantic locations, and $\hat{T}$ be the summarized semantic trajectory derived from the former. Moreover, let $R(\cdot)$ be the number of distinct locations in a trajectory. Then, we define the summarization rate achieved by a summarized semantic trajectory $\hat{T}$ as: $S_{rate}(\hat{T}) = 1 - (R(\hat{T})/R(\tilde{T}))$. S_{rate} values close to 0 imply poor summarization, while values close to 1 indicate a high degree of summarization. It is important to recall that in $\tilde{T}$ each point p_i is seen as a distinct location. Conversely, in a summarized semantic trajectory we treat each representative semantic location as a distinct location. Finally, the overall summarization rate for the entire dataset is obtained by computing the average of the values of the summarization rate for each trajectory.

Example. Consider the toy trajectory $\tilde{T}$ introduced in Section 4.2: $\tilde{T} = \langle (p_1, A, 1), (p_2, A, 1)(p_3, B, 1), (p_4, C, 1)(p_5, B, 1)(p_6, B, 1)\rangle$, where $R(\tilde{T}) = 6$. Assuming the summarized semantic trajectory is $\hat{T} = \{(\overline{A}, 2)(\overline{B}, 4)\}$, resulting in $R(\hat{T}) = 2$, we find that $S_{rate}(\hat{T}) = \frac{2}{3}$. We instantiate the semantic quality measure using the *MUITAS* similarity metric [19]. This metric can compare trajectories of different lengths. *MUITAS* scores close to 1 mean a strong semantic similarity between two trajectories, while scores close to 0 indicate a strong dissimilarity. Within *MUITAS*, each semantic aspect is paired with a corresponding distance function. Additionally, *MUITAS* requires two more parameters for each aspect: a maximum distance threshold and a weight. Then, for each pair of trajectory points, *MUITAS* applies each distance function to the instances of its aspect found in the pair of points: if the result falls below the distance threshold, there is a match. The match is then weighted for each aspect according to the associated weight. The average of these matches results in a maximum score of 1 when the two trajectories are semantically identical and 0 when they are entirely dissimilar. It is then clear that *MUITAS* can be used as an indicator of the summarized trajectories' ability to preserve semantic quality effectively. Finally, we determine the score of *MUITAS* over an entire dataset of summarized trajectories by averaging the *MUITAS* scores between each weighted sequence of semantic locations and its corresponding summarized counterpart.

It is important to highlight that we define what is called in [19] the feature set, i.e., the set of semantic aspects, by considering both the original trajectory's semantic aspects (if any) and the aspects of the semantic locations. This approach enables evaluating the

[1] https://www.openstreetmap.org/
[2] https://wiki.openstreetmap.org/wiki/Map_features
[3] https://github.com/chiarap2/MAT-Sum

potential loss of semantics resulting from the trajectory summarization step. In the Foursquare dataset the original trajectories are already enriched with several aspects, i.e., POI category, rating, price, weather, and weekday, while in the Geolife dataset, the trajectories are already enriched with the means of transportation. We also recall that the tiles are geographically enriched with aspects such as POIs, land use, and public transportation. Following the suggestion in [19], we use the Euclidean distance function to compare continuous values (e.g., price and rating), binary distance for discrete values (e.g., means of transportation or weather), and check whether one set is a subset of the other when comparing two sets corresponding to the semantic context of cells and locations. We assign each semantic aspect the same weight since, for the purposes of our experiments, each aspect has the same importance. Finally, the similarity between semantic locations (see Section 4.2, Equation 2) is implemented with the *cosine similarity*.

5.2 Experimental Results

We conduct an experimental evaluation of MAT-Sum by addressing the research question RQ2. To this end, we rephrase RQ2 into three distinct experimental questions:

- **EQ1** To what extent does the MAT-Sum approach prove effective in summarizing trajectories?
- **EQ2** What is the level of semantic quality achieved by MAT-Sum?
- **EQ3** How do the results of the MAT-Sum approach compare with the baselines?

To deliver an exhaustive evaluation of our method, we vary several parameters involved in its execution. Specifically, we examine multiple scenarios that include: (1) variations in the tessellation method, from uniform squares with resolution (i.e., zoom level that indicates how zoomed-in the tile is) 16, 17, or 18, to hexagons with resolution 6, 7, or 8; (2) changes in the number of semantic aspects associated with the tiles like POIs, land use, and public transportation; (3) variations of the semantic locations similarity threshold τ between 0.6 and 0.9. The experimental results, conducted on both datasets, are reported in Tables 1 and 2. In Table 1, we consider all the semantic aspects used to semantically enrich the cells (i.e., POIs, land use, and public transport), a value of $\tau = 0.9$, and the effect of different tessellation methods as well as the relative resolution on the summarization rate and semantic quality. In Table 2, we study the variation of the summarization rate and semantic quality when considering different sets of semantic aspects and τ thresholds, adopting the square tessellation with resolution 17.

EQ1: To what extent does the MAT-Sum approach prove effective in summarizing trajectories? Table 1 reveals that MAT-Sum achieves lower values of S_{rate} when applying the uniform grid tessellation; this is particularly noticeable with the Foursquare dataset. Conversely, S_{rate} values increase with hexagonal tessellation. However, for the Geolife dataset, performance remains consistently high irrespective of the chosen tessellation. This can be due to the significant discrepancy between the length of GPS trajectories (in terms of samples) and the size of semantically summarized trajectories.

Inspection of Table 2 shows the highest S_{rate} occurring when solely considering the land use aspect. This suggests that trajectories enriched with this aspect can be more effectively summarized than those with other semantic aspects. In contrast, when taking the POI aspect into account, we record the lowest S_{rate}. This suggests that tiles enriched with this aspect contain more diverse and distinct information, complicating trajectory summarization. When considering the public transport semantic aspect, MAT-Sum achieves average results in terms of summarization rate. The differences between the three aspects are particularly evident when analyzing the Foursquare dataset.

		Foursquare NYC		Geolife	
Tessellation	**Resolution**	S_{rate}	**MUITAS**	S_{rate}	**MUITAS**
Squares	16	0.5536	0.8662	0.9531	0.5135
	17	0.4680	0.8770	0.9383	0.5999
	18	0.4413	**0.8987**	0.9418	**0.7411**
Hexagons	6	**0.8752**	0.8571	**0.9894**	0.4915
	7	0.7866	0.8571	0.9765	0.4984
	8	0.6532	0.8600	0.9637	0.4981

Table 1: Results of S_{rate} and *MUITAS* metrics, varying tessellation methods and resolutions, with fixed all semantic aspects and $\tau = 0.9$.

		Foursquare NYC		Geolife	
Semantic aspects	τ	S_{rate}	**MUITAS**	S_{rate}	**MUITAS**
Land use	0.7	0.8813	0.9991	0.9827	0.9630
	0.8	0.8174	0.9980	0.9800	0.9643
	0.9	0.7211	0.9976	0.9726	0.9611
Public transport	0.7	0.8096	0.8739	0.9857	0.6257
	0.8	0.7641	0.8724	0.9832	0.6125
	0.9	0.7156	0.8721	0.9782	0.6125
POIs	0.7	0.7491	0.8773	0.9739	0.6661
	0.8	0.6577	0.8727	0.9609	0.5488
	0.9	0.5509	0.8713	0.9545	0.5442

Table 2: Results of S_{rate} and *MUITAS* metrics, varying the semantic aspects used to enrich tiles and τ, with fixed tessellation method (uniform grid) and resolution (17).

Overall, the results align with our initial expectations: the number and variety of semantic aspects used for geographical enrichment significantly impact the summarization process. As we increase the number of instances and aspects, the resulting semantic locations become more unique and consequently harder to summarize effectively. This observation aligns with the intuition that an increased level of semantic richness and diversity in the data entails greater complexity in the trajectory summarization task. Consequently, striking a balance between effective summarization and preservation of semantic details becomes increasingly challenging in such scenarios.

EQ2: What is the level of semantic quality achieved by MAT-Sum? Table 1 shows that, particularly in the case of Geolife, the *MUITAS* score is higher when employing a square tessellation instead of a hexagonal one, which indicates better preservation of semantics. An analysis of the results presented in Table 2 reveals that as we consider an increasing number of semantic aspects for

tile enrichment, the *MUITAS* score decreases for both datasets. However, a notable distinction exists between the two datasets. Although the *MUITAS* score for Foursquare remains relatively high, even when increasing the number of considered aspects, the same trend does not hold for Geolife. This discrepancy can be ascribed to two factors: (1) the substantial disparity in length between semantic and summarized trajectories, and (2) the original Geolife dataset contains significantly fewer semantic aspects than Foursquare.

To sum up, MAT-Sum achieves good semantic quality when applied to the Foursquare dataset, even considering all semantic aspects and a high τ value. Conversely, when applied to the Geolife dataset, MAT-Sum retains a certain level of semantics, although it encounters some difficulties when comparing trajectories of different lengths.

EQ3: How do the results of the MAT-Sum approach compare with the baselines? Finally, we compare our results against the two baselines introduced in Section 5.1, namely, RLE and Seqscan-D. In particular, we compare MAT-Sum with RLE on both datasets and with Seqscan-D only on Geolife. Indeed, Seqscan-D cannot be applied to the Foursquare dataset for two reasons: it is not designed to manage sparse trajectories such as those found in Foursquare, and it chooses which locations to eliminate based on the time spent in each of them, a piece of information not available in Foursquare.

In Figure 3, we adopt the same parameter configurations used in Tables 1 and 2. The comparison between MAT-Sum and RLE on the Foursquare dataset is shown in Figures 3a and 3b. Additionally, the Figures 3c and 3d illustrate the comparison between MAT-Sum and both baselines on the Geolife dataset. Note that when only one color is visible per bar in the plots, this indicates equivalent values across all bars. From Figure 3, we observe that our method always outperforms RLE in terms of S_{rate}. If we fix $aspects = all$ and $\tau = 0.9$, the results obtained with RLE with the uniform square tessellation are very close to those obtained with MAT-Sum. However, our method widens the gap with RLE when the hexagonal tessellation is used, as shown in the dashed part of Figure 3a. This difference becomes even more apparent when considering the results on the Geolife dataset, as illustrated in Figure 3c. Additionally, Figure 3c reveals that Seqscan-D performs slightly better than MAT-Sum.

When we vary the number of semantic aspects and τ while keeping the tessellation method and resolution fixed, we consistently observe similar behaviors as delineated in the preceding experiments. Specifically, MAT-Sum consistently achieves higher S_{rate} values compared to RLE on the Foursquare dataset (see Figure 3b). This difference becomes even more evident when analyzing the Geolife dataset (see Figure 3d). Similarly, when comparing MAT-Sum to Seqscan-D under the same parameter configuration, we observe that Seqscan-D slightly outperforms MAT-Sum in terms of S_{rate}.

These results are in line with our expectations. RLE aggregates consecutive equal labels, resulting in lower S_{rate} values – this is particularly evident in the Geolife case. In contrast, Seqscan-D selectively discards irrelevant locations, leading to a stronger summarization than our method (which maintains all locations). To address this disparity, we evaluated the summarized trajectories of both approaches using *MUITAS* to demonstrate that MAT-Sum preserves more semantic information from the original trajectory than Seqscan-D. The results, shown in Table 3, highlight that MAT-Sum presents a higher *MUITAS* score than Seqscan-D for all the configurations. Thus, considering that the goals of the two baseline methods slightly differ from our approach, it is worth noting that MAT-Sum outperforms RLE in terms of S_{rate}, particularly with the Geolife dataset and hexagonal tessellation. Compared to Seqscan-D, although MAT-Sum achieves slightly lower S_{rate}, it compensates for this by preserving higher semantic quality.

Tessellation	Resolution	MUITAS (MAT-Sum)	MUITAS (Seqscan-D)
Squares	16	0.5135	0.0129
	17	0.5999	0.0056
	18	**0.7411**	0.002
Hexagons	6	0.4915	**0.2299**
	7	0.4984	0.0753
	8	0.4981	0.0417

Table 3: Comparison between Seqscan-D and MAT-Sum in terms of MUITAS metric.

6 CONCLUSIONS AND FUTURE WORKS

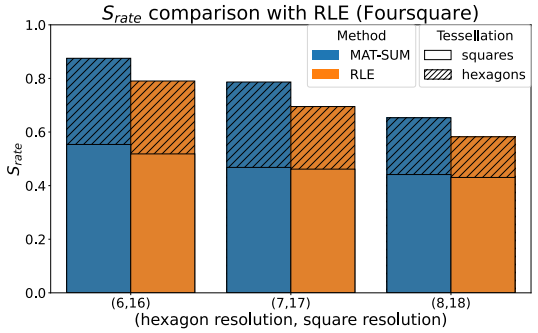
In this paper, we present a novel method, named MAT-Sum, which summarizes trajectories using semantically enriched geographical contexts. This overcomes issues associated with existing approaches in handling the complexity of heterogeneous semantic dimensions and the massive size of movement data. MAT-Sum leverages a location-centric enrichment of the trajectories, maximizing the summarization rate while preserving a good level of semantic quality. We assessed the effectiveness of MAT-Sum across various scenarios, including (1) different tessellation methods and tile sizes, (2) variations in the number of associated semantic aspects for each tile, and (3) adjustments to the semantic location similarity threshold, τ. Additionally, we compared our method with two baseline approaches that address trajectory summarization. In general, MAT-Sum exhibits good performance in both summarization rate and preservation of semantic quality, as corroborated by our comparative experiments. It either surpasses the competing approaches in terms of summarization rate or preserves more semantic information. Looking forward, we plan to expand our work in several directions. First, we aim to enhance the quality of semantics considered in each trajectory by incorporating the temporal evolution of the geographical context. Then, we intend to extend MAT-Sum to retain the sequential information of the trajectories and study its impact on the effectiveness of the method. In conclusion, we aim to investigate other semantic aspects to enrich tiles and analyze the resultant summarized trajectories.

Acknowledgements. This work has been supported by the EC H2020 projects MOBIDATALAB (GA 101006879), MASTER (GA 777695), and SoBigData++ (GA 871042), SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

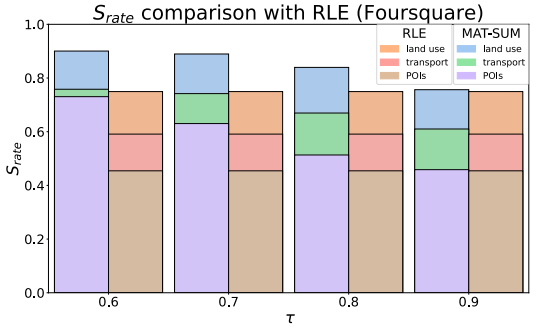
REFERENCES

[1] D. Amigo, D. S. Pedroche, J. García, and J. M. Molina. Review and classification of trajectory summarisation algorithms: From compression to segmentation. *Int. J. of Distributed Sensor Networks*, 17(10):15501477211050729, 2021.

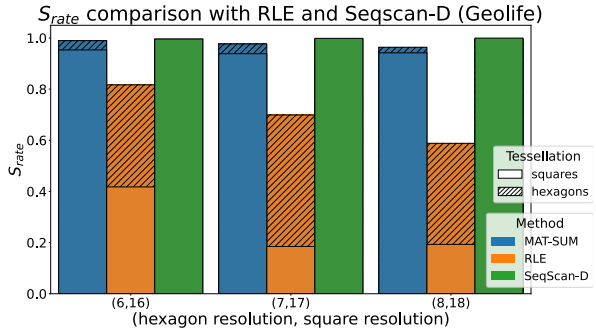
[2] V. Bogorny, C. Renso, A. R. de Aquino, F. de Lucca Siqueira, and L. O. Alvares. CONSTAnT - A conceptual data model for semantic trajectories of moving objects. *Trans. GIS*, 18(1):66–88, 2014.



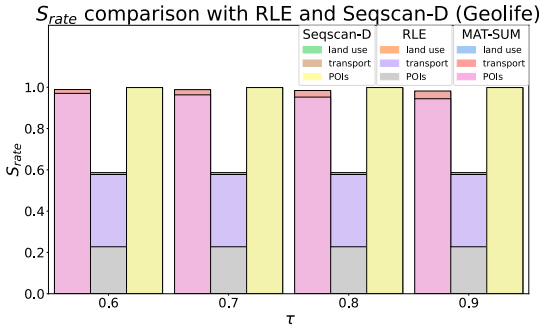
(a) All semantic aspects considered and $\tau = 0.9$



(b) Fixed tessellation (uniform grid) and resolution (17)



(c) All semantic aspects considered and $\tau = 0.9$



(d) Fixed tessellation (uniform grid) and resolution (17)

Figure 3: MAT-Sum results compared with two baselines (RLE and Seqscan-D) in terms of S_{rate}.

[3] V. Chandola and V. Kumar. Summarization - compressing data into an informative representation. *Knowl. Inf. Syst.*, 12(3):355–378, 2007.

[4] Y. Chen, K. Jiang, Y. Zheng, C. Li, and N. Yu. Trajectory simplification method for location-based social networking services. In *Proceedings of the 2009 international workshop on location based social networks*, pages 33–40, 2009.

[5] M. L. Damiani and F. Hachem. Segmentation techniques for the summarization of individual mobility data. *WIREs Data Mining Knowl. Discov.*, 7(6), 2017.

[6] M. L. Damiani, F. Hachem, C. Quadri, M. Rossini, and S. Gaito. On location relevance and diversity in human mobility data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 7(2):1–38, 2020.

[7] R. Fileto, C. May, C. Renso, N. Pelekis, D. Klein, and Y. Theodoridis. The baquara^2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data Knowl. Eng.*, 98:104–122, 2015.

[8] C. Fu, H. Huang, and R. Weibel. Adaptive simplification of GPS trajectories with geographic context - a quadtree-based approach. *Int. J. Geogr. Inf. Sci.*, 35(4):661–688, 2021.

[9] R. H. Güting, F. Valdés, and M. L. Damiani. Symbolic trajectories. *ACM Trans. Spatial Algorithms Syst.*, 1(2), jul 2015.

[10] T. Hagen, J. Hamann, and S. Saki. *Discretization of Urban Areas Using POI-based Tesselation.* Working papers. Frankfurt University of Applied Sciences, Fachbereich 3: Wirtschaft und Recht, 2022.

[11] Z. R. Hesabi, Z. Tari, A. M. Goscinski, A. Fahad, I. Khalil, and C. Queiroz. Data summarization techniques for big data - A survey. In S. U. Khan and A. Y. Zomaya, editors, *Handbook on Data Centers*, pages 1109–1152. Springer, 2015.

[12] F. Lettich, C. Pugliese, C. Renso, and F. Pinelli. General methodology for building multiple aspect trajectories. In *The 38th ACM/SIGAPP Symposium On Applied Computing, ACM SAC 2023, Tallin, Estonia, March 27-31, 2023, Proceedings*, 2023.

[13] M. Liu, G. He, and Y. Long. A semantics-based trajectory segmentation simplification method. *J. of Geo. and Spatial Analysis*, 5:1–15, 2021.

[14] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo. A survey on deep learning for human mobility. *ACM Comput. Surv.*, 55(1), nov 2021.

[15] V. L. Machado, R. dos Santos Mello, and V. Bogorny. A method for summarizing trajectories with multiple aspects. In *DEXA 2022, Vienna, Austria, August 22-24*, volume 13426 of *Lecture Notes in Computer Science*, pages 433–446. Springer, 2022.

[16] R. d. S. Mello, V. Bogorny, L. O. Alvares, L. H. Z. Santana, C. A. Ferrero, A. A. Frozza, G. A. Schreiner, and C. Renso. MASTER: A multiple aspect view on trajectories. *Transactions in GIS*, 23(4):805–822, 2019.

[17] T. P. Nogueira, R. B. Braga, C. T. de Oliveira, and H. Martin. Framestep: A framework for annotating semantic trajectories based on episodes. *Expert Systems with Applications*, 92:533–545, 2018.

[18] A. Okabe, B. Boots, K. Sugihara, and S. Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, volume 43. John Wiley & Sons, 05 2000.

[19] L. M. Petry, C. A. Ferrero, L. O. Alvares, C. Renso, and V. Bogorny. Towards semantic-aware multiple-aspect trajectory similarity measuring. *Transactions in GIS*, 23(5):960–975, 2019.

[20] C. Pugliese, F. Lettich, C. Renso, and F. Pinelli. MAT-builder: a system to build semantically enriched trajectories. In *MDM 2022*, pages 274–277, 2022.

[21] L. Ruback, M. A. Casanova, A. Raffaetà, C. Renso, and V. M. P. Vidal. Enriching mobility data with linked open data. In *IDEAS 2016*, pages 173–182. ACM, 2016.

[22] K. Sahr. Central place indexing: Optimal location representation for digital earth using hierarchically indexed mixed-aperture hexagonal discrete global grids. In *AutoCarto 2014, the 20th International Research Symposium on Computer-based Cartography, Pittsburgh, Pennsylvania, USA, October 5-7, 2014.*, 2014.

[23] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot. A conceptual view on trajectories. *DKE*, 65(1):126–146, 2008.

[24] R. Tamilmani and E. Stefanakis. Modelling and analysis of semantically enriched simplified trajectories using graph databases. *Advances in Cartography and GIScience of the ICA*, 1:20, 2019.

[25] D. Teixeira, J. Almeida, and A. Viana. On estimating the predictability of human mobility: the role of routine. *EPJ Data Science*, 10, 12 2021.

[26] S. Wang, Z. Bao, J. S. Culpepper, and G. Cong. A survey on trajectory data management, analytics, and learning. *ACM Comput. Surv.*, 54(2), mar 2021.

[27] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2015.

[28] Y. Zheng, X. Xie, and W.-Y. Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33:32–39, 2010.