Journal of Information and Data Management, 2025, 16:1, doi: 10.5753/jidm.2025.4110

Towards Data Summarization of Multi-Aspect Trajectories Based on Spatio-Temporal Segmentation

Vanessa Lago Machado [**Universidade Federal de Santa Catarina (UFSC), Instituto Federal Sul-Rio-Grandense (IFSUL)** | ***vanessalagomachado@gmail.com***]
Tarlis Tortelli Portela [**Instituto Federal do Paraná (IFPR)** | ***tarlis@tarlis.com.br***]
Geomar André Schreiner [**Universidade Federal da Fronteira Sul (UFFS)** | ***gschreiner@uffs.edu.br***]
Ronaldo dos Santos Mello [**Universidade Federal de Santa Catarina (UFSC)** | ***r.mello@ufsc.br***]

Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, Bairro Trindade, Florianópolis, SC, 88040-900, Brasil.

Received: 7 February 2024 • **Published:** 14 January 2025

Abstract This paper presents a new method for summarizing multiple aspect trajectories (MATs). This kind of data holds several challenges in terms of analysis and extraction of meaningful insights due to their spatial, temporal, and semantic dimensions. In order to address them, our method leverages a combination of spatial grid-based segmentation and temporal sequence analysis. It segments the trajectory data into spatial cells using a grid-based approach. The spatial segmentation enables a finer-grained analysis of the trajectories within each cell. Next, we consider the temporal sequence of points within each cell to capture the temporal intervals of the trajectories. By combining spatial and temporal perspectives, the method identifies representative trajectories that capture the main behavior of semantically enriched object movements. We evaluated the utility of our method by applying two distinct strategies: (i) the RMMAT measure, assessing the quality of representative MAT in terms of similarity and coverage of information, and (ii) the Average Recall (AR) metric, measuring the ability of our representative MAT to capture essential data characteristics. Our evaluation demonstrates the effectiveness of MAT-SGT in summarizing MATs. The proposed method holds potential applications across diverse domains, including transportation planning, urban analytics, and human mobility analysis, where the concise representation of trajectories is crucial for decision-making and knowledge discovery.

Keywords: Multiple aspect trajectory, representative trajectory, trajectory summarization

1 Introduction

Analyzing mobility data and understanding movement patterns is crucial for various purposes. It helps analyze the movements of different objects, like people, vessels, and animals. Consequently, this analysis could help understand patterns like animal migration and natural phenomena like hurricanes. Typically, these mobility data are represented as a sequence of points with spatial and temporal information *(x, y, t)*, known as *raw trajectories* [Erwig *et al.*, 1999].

When a raw trajectory is enriched with semantic information, such as points of interest (PoIs) visited by the moving object, these trajectories are known as *semantic trajectory*. When a trajectory or its points are associated with multiple semantic contexts, it is referred to as a *Multiple Aspect Trajectory (MAT)* [Mello *et al.*, 2019].

Trajectory data is often generated continuously and frequently, requiring efficient storage and processing to avoid overwhelming computing systems. MAT data comprises three dimensions (*spatial*, *temporal*, and *semantics*), with the third one potentially holding a lot of aspects, providing a large data volume that could have vast heterogeneity. For example, the spatial position of a point in a specific timestamp can be associated with a *PoI* with a *name* and a *category* (e.g., Hotel, School, Restaurant). Depending on the type, specific attributes may hold, such as *price* and *rating* for a hotel.

The MAT provides a complex representation of information about moving objects. However, this complexity poses a challenge for trajectory data mining since extracting meaningful insights from the voluminous and complex MAT data is a critical task. Innovative approaches are needed to achieve this task successfully, as it is crucial for practical analysis, informed decision-making, and solving complex mobility patterns.

In response to these challenges, trajectory summarization methods have emerged as invaluable tools to distill essential information from these massive datasets, aiming to reduce this complexity. By computing representative trajectories from a set of data, these data can be used to teach recommendation systems about individual movement patterns, for example, which can then be utilized to provide personalized suggestions based on user preferences and behaviors. While surveys have been addressing trajectory data, its summarization of semantic information remains an open issue [Fiore *et al.*, 2020; Wang *et al.*, 2021]. This lack of research is probably due to the inherent complexity of these data, as different semantic contexts may coexist and be related to various parts of a trajectory, making data summarization tasks more challenging. The main challenge regarding MATs summarization is reducing data volume and variety by computing *representative data*, allowing the discovery of the most relevant information. Additionally, the effectiveness of calculating a representative trajectory depends on its intended use.

Prior research mainly focused on reducing raw trajectory data, emphasizing spatial dimension [Buchin *et al.*, 2013, 2019; Etienne *et al.*, 2016; Gao *et al.*, 2019; Lee *et al.*, 2007]. While recent studies have delved into extracting representative data from MATs [Seep and Vahrenhold, 2019; Machado *et al.*, 2022], there remains a gap in encompassing data representing both spatial and temporal movement sequences, summarizing all aspects involved in the original data.

Then, our proposed method, *MAT-SGT*, emerges as a promising solution for summarizing MATs. By leveraging a spatial grid and temporal intervals, *MAT-SGT* strategically identifies and distills temporal intervals as the key information, capturing the main behaviors and features inherent in the input MATs. Data volume reduction is achieved with minimal loss of utility, addressing the core challenges associated with MATs summarization. To provide a comprehensive understanding, we delve into detailed comparisons with related work in Section 3, shedding light on the contributions of *MAT-SGT* in the landscape of trajectory summarization.

Additionally, this paper refers to an extended version of Machado *et al.* [2023a], presented at XXIV Brazilian Symposium on GeoInformatics (GEOINFO 2023). We have significantly improved Section 2 by adding more conceptual information about trajectory summarization and representative trajectory. To clarify further, we added a Problem Definition section (Section 4).

In addition, we have expanded Section 5 to provide more detailed information about the method, including improvements in architecture details and specifics about the output data. Furthermore, we have extended Section 6 to include a comparative experimental evaluation with the state-of-the-art MAT summarization method *MAT-SG*. Through this comparative analysis, we aim to assess the effectiveness of *MAT-SGT* and gain insights into representative data computation. This analysis will also aid in understanding the differences between various summarization methods.

We evaluate our approach using two distinct strategies: (i) the RMMAT measure, evaluating the quality of representative MAT in terms of both similarity and coverage of information, and (ii) the Average Recall (AR) metric, aiming to measure the quality of our representative MAT in capturing essential data characteristics.

The RMMAT measure provides a comprehensive evaluation by quantifying how well the *RT* represents the underlying trajectory data through its similarity and coverage. This measure is particularly important as it offers a nuanced understanding of the effectiveness of our proposed summarization method, *MAT-SGT*. By employing these two evaluation strategies, we aim to demonstrate the robustness of our approach in summarizing multiple aspect trajectories.

To broaden the scope of our experiments, we included three datasets of user trajectory data: Foursquare (193 users), Gowalla (300 users), and Brightkite (300 users). The Foursquare dataset consists of check-in data from New York City, offering rich spatial, temporal, and semantic insights into user interactions with various PoIs. The Gowalla dataset is collected globally, providing valuable insights into social interactions and mobility patterns across different locations and times. The Brightkite dataset refers to trajectory data from a social media platform, enabling the analysis of user movement patterns and behaviors in urban environments over a specified time period.

The remainder of this paper is organized as follows. Section 2 presents the basic concepts associated with *MAT-SGT*. Section 3 is dedicated to related work. Section 4 defines the problem. Section 5 describes the proposed method. Section 6 presents the evaluation, and Section 7 concludes the paper, outlining avenues for future work.

2 Basic Concepts

Trajectory data, as stated in the previous section, captures the sequential movement of objects in space and time. The increasing availability of *Location-Based Services (LBS)* and sensor technologies has led to voluminous and complex trajectory data, giving rise to MATs [Mello *et al.*, 2019]. MATs capture the sequential movement of objects and encompass various aspects that reflect object movement behavior and characteristics.

Definition 2.1 (**Multiple Aspect Trajectory**). *A MAT is a sequence of points* $(p_1, p_2, ..., p_n)$*, with* p_i *= (x, y, t, A) being the i-th point of the trajectory generated in the location (x,y) at timestamp t, and described by the set* $A = \{a_1 : v_1, a_2 : v_2, ..., a_r : v_r\}$ *of r aspect-value pairs that characterize various aspects of the trajectory.*

In short, an *aspect* represents relevant real-world facts such as social media posts, weather conditions, or transportation modes. Each aspect a_i is characterized by attributes that provide detailed information about the aspect. By encompassing multiple aspects, MAT enables a more comprehensive understanding of the underlying trajectory data.

Figure 1 illustrates a MAT of an individual over one day. It includes diverse information such as transportation modes, social media postings, weather conditions, and health information. As emphasized, the initial segment of the trajectory, between 11 pm and 8 am, consists of a set of data points in the same location. Each data point includes critical aspects: geographical coordinates, timestamps, and semantic aspects such as PoI (”Home”) information and health information such as heart rate and sleep stages. This example highlights the complexity of MATs, as they comprise attributes from multiple heterogeneous aspects, making the analysis and extraction of meaningful insights challenging.

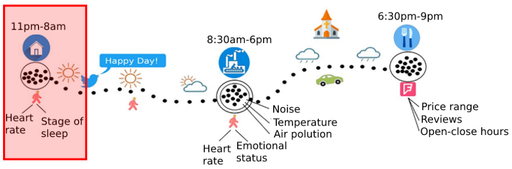


Figure 1. A MAT describing an individual movement (Adapted from Mello *et al.* [2019]).

Summarization methods are techniques used to condense information, which makes it easier to analyze large volumes of data [Ermakova *et al.*, 2019]. Trajectory summarization, in turn, refers to reducing the volume of trajectory data while preserving its essential characteristics and patterns. By summarizing trajectories, we can achieve a more compact representation that retains relevant information [Hesabi *et al.*, 2015].

While trajectory compression also results in compact data, it represents a distinct process from summarization. Compression typically involves two components: *(i)*, an encoding algorithm that transforms the original data into a compressed format, and *(ii)*, a decoding algorithm that reconstructs the original data or an approximation thereof from the compressed representation. Although compression yields compact data, it often results in unintelligible representations. In contrast, summarization provides an intelligible representation of the data, which enhances further analysis and decision-making capabilities [Machado *et al.*, 2024].

Representative trajectories provide a concise and informative presentation of the input dataset, facilitating analysis, visualization, and other tasks based on the trajectories [Machado *et al.*, 2022].

According to Lee *et al.* [2007], a representative trajectory is an imaginary trajectory that denotes the main behavior of a cluster of trajectories. As noted in Panagiotakis *et al.* [2009], the definition of a representative trajectory can vary according to the focus of interest, such as density, frequency, and pairwise distance.

Lets $D = \{T_1, T_2, ..., T_n\}$ a set of n trajectories, formally the representative trajectory data *RT* is defined as follows.

Definition 2.2 (Representative trajectory). *A representative trajectory is a compact and informative abstraction that aims to balance data reduction with the retention of essential features. The purpose of computing (RT) is to capture the main behaviors and patterns from the original dataset (D) while minimizing the loss of critical information. The resulting trajectory (RT) should represent the most frequent patterns or characteristic movements present in the dataset.*

The process of computing (*RT*) involves the following steps, in line with the approaches outlined in the literature [Machado *et al.*, 2024]:

1. *Clustering or segmenting* similar trajectories based on spatial, temporal, or semantic dimensions.
2. Identifying a *centroid trajectory* or selecting a set of key points within each cluster that best represents the overall movement pattern.
3. Quantifying the resemblance between the original trajectories and the representative trajectory.

It is important to note that the *representative trajectory* is not necessarily a real trajectory; rather, it is an *imaginary or synthesized trajectory* that captures the *main behaviors* of the data [Machado *et al.*, 2024]. It provides a compact yet representative summary of the trajectory dataset.

In summary, employing representative data to understand the patterns within a set of MATs offers a powerful solution to tackle the challenges arising from the volume and complexity of trajectory data, enabling more efficient storage, processing, and analysis. It is important to note that the effectiveness of trajectory data summarization depends on the specific purpose for which the representative data is intended. Different applications or analysis tasks may require different levels of granularity and information preservation [Ahmed, 2019]. Therefore, the computation of *RT* should align with the specific objectives and requirements of the intended use case.

3 Related Works

In recent years, the analysis of MATs has gained significant attention due to the increasing availability of location-based data. Various methods have been proposed to summarize and analyze these complex datasets, each with its strengths and limitations. The task of computing representative data that balances quality and utility is challenging. The ultimate goal is to ensure that the representative data retains sufficient information about the original trajectories while minimizing data loss. Previous research on trajectory data reduction and summarization has predominantly focused on raw trajectory data [Buchin *et al.*, 2013, 2019; Etienne *et al.*, 2016; Gao *et al.*, 2019; Lee *et al.*, 2007], acknowledging that MATs present specific challenges requiring specialized treatment [Mello *et al.*, 2019].

The effectiveness of computing a representative trajectory depends on its specific purpose. MATs differ from simple trajectories by incorporating semantic data that enriches the spatial and temporal dimensions, adding crucial context. For instance, in user mobility, the spatial dimension captures locations visited, the temporal dimension tracks time spent at each location, and the semantic dimension provides the purpose of visits and contextual factors such as traffic conditions or weather. Similarly, for public transportation vehicles, the spatial trajectory follows roads and makes stops, the temporal aspect provides timing information, and the semantic aspect includes details like the type of vehicle, route number, or passenger load. Integrating these semantic aspects into MAT summarization helps distinguish between different types of behavior, leading to higher-quality summarization that provides deeper insights into the data.

In a recent paper, Pugliese *et al.* [2023] presented a novel approach for MAT summarization that computes representative data based on a set of raw trajectories, where they are enriched with semantic context. However, this paper creates group representative data for each group.

In this paper, we focus on methods that generate a single representative data by summarizing a group of MATs. One of the early works in this direction was presented by Seep and Vahrenhold [2019], which treats all attributes of the MAT points as spatial or non-spatial data without considering the individual analysis of semantic data as categorical or numerical. The method utilizes a Finite State Machine (FSM) to identify a sequence of common transitions among the movements, where each state represents a common point, and a sequence of states yields the representative trajectory. However, it is important to note that this work lacks sufficient detailed information, as it is a short paper, making it hard to understand and fully reproduce the method.

In 2021, a closely related approach to trajectory data summarization was introduced by Varlamis *et al.* [2021], which presents navigation networks derived from multi-vessel trajectory data. This builds upon their earlier work, which proposed a network abstraction model to detect anomalies in maritime traffic using multi-vessel trajectories Varlamis *et al.* [2019]. Their approach provides a structural summary of vessel movements to detect anomalies in maritime traffic by leveraging spatial and temporal dimensions and vessel velocity. However, the approach does not fully address the complexity of MATs without integrating additional semantic aspects (e.g., specific purposes or activities associated with those movements).

More recently, in 2022, the *MAT-SG* method [Machado *et al.*, 2022] was proposed as a comprehensive data summarization method for MATs. Unlike the previous method, *MAT-SG* treats all aspects of data individually, enabling the identification of patterns and the understanding of the influence of each aspect on the representative trajectory. It also defines mappings between input MATs and the representative data.

However, the *MAT-SG* method comprises spatial segmentation and data summarization. Initially, the input MATs are segmented into spatial cell grids, and then data summarization is performed within each cell. Consequently, the representative trajectory reflects the patterns specific to each spatial area. While *MAT-SG* addresses various dimensions and treats each semantic type individually, it lacks the identification of temporal sequences within the movement patterns. In contrast, our novel method, *MAT-SGT*, is a straightforward data summarization method specifically designed to compute representative MATs identifying the temporal sequence associated with the movement pattern. At the same time, it includes mappings between input MATs and the representative MAT.

Table 1 provides a comparison of methods for MAT summarization in terms of the aspects considered in the movement pattern. The *Aspects Considered* column indicates whether each dimension of the MAT is completely (✓) or partially (□)*consideredbythesummarizationprocess.TheMovement Patterncolumnsuggeststhesameaboutthedimensionsinvolvedinthem*

Table 1. Related work comparison

Method X MAT Summarization Analysis		[Seep and Vahrenhold, 2019]	Varlamis *et al.* [2021]	*MAT-SG*	*MAT-SGT*
Aspects Considered	**Spatial**	✓	✓	✓	✓
	Time	□	✓	✓	✓
	Semantic	□	□	✓	✓
Movement Pattern	**Spatial**	✓	✓	✓	✓
	Time		✓		✓
	Semantic		□		
Mapping Information			✓	✓	✓

As observed, several studies have focused on spatial segmentation techniques to enhance trajectory analysis. Temporal analysis is another critical aspect of trajectory summarization. Previous works, such as Varlamis *et al.* [2021], explored temporal sequence analysis in the context of vessel trajectories by abstracting them into a network model. This gap emphasizes the necessity of a more comprehensive approach that takes into account both temporal sequences and semantic enrichment, which is the focus of our method, *MAT-SGT*. The incorporation of semantic information into trajectory analysis has been explored in various studies, such as Mello *et al.* [2019], which demonstrated the benefits of semantic enrichment for understanding user behavior. However, most of these methods focus on spatial or temporal dimensions, lacking a comprehensive framework combining all three aspects. Our research fills this gap by proposing a method that simultaneously addresses spatial, temporal, and semantic dimensions, offering a more complete MAT summarization.

While existing methods have made significant contributions to the field, they often exhibit limitations in capturing the full complexity of MATs. For instance, the FSM-based approach by Seep and Vahrenhold [2019] struggles with spatial representation, and the *MAT-SG* focuses primarily on spatial areas without comprehensively addressing temporal sequences. Similarly, Varlamis *et al.* [2021] excel in identifying representative trajectories through a network abstraction model that includes spatial and temporal information but only considers one semantic aspect (vessel velocity). The limited use of semantic dimensions in this network model reduces its ability to provide context-aware insights, leaving a more nuanced understanding of the complexity of MATs as an open issue. Our proposed *MAT-SGT* overcomes these limitations by simultaneously incorporating spatial, temporal, and multiple semantic dimensions to provide a richer, more comprehensive MAT summarization.

Additionally, while *MAT-SGT* builds on the foundation of *MAT-SG*, its integration of the temporal sequence analysis allows it to handle more complex tasks and scenarios. For example, consider the task of summarizing customer movements in a shopping mall. *MAT-SG* can summarize spatial paths through the mall, showing that certain areas (e.g., food courts or entrances) are frequently visited and associating these locations with contextual information, such as what time periods these areas are busy, on which days this occurs, and under what conditions (e.g., weather or promotions). However, *MAT-SG* is unable to capture how customer behavior changes throughout the day.

For instance, customers might visit the food court during lunch hours and then proceed to the supermarket, while other stores may be more popular in the evening after customers have eaten. *MAT-SGT* addresses this limitation by incorporating the temporal sequence of movement behavior, tracking both the sequence of places visited and the sequence of corresponding time intervals. This temporal sequence insight enables mall managers to optimize staffing and marketing strategies based on time-based customer behavior patterns—providing a level of detail that *MAT-SG* cannot achieve.

4 Problem Definition

The computation of representative information in trajectory data should be aligned with specific use case objectives and requirements since different applications may require different levels of granularity and information preservation [Machado *et al.*, 2022]. This paper intends to answer this fundamental question: 'How can we effectively summarize a set of input MATs to compute representative data that captures and reflects the essence of the original MATs within an input dataset **T**, while also providing the temporal sequence of the data?'.

To understand the pattern in data and the temporal sequence of data, we have focused on the following preliminaries in this paper. First, trajectory data primarily comprises a sequence of points with spatial and temporal information. In MAT, we can have many semantic aspects involved with these data points. Then, it is essential to consider that (i) the spatial dimension is crucial for understanding and preserving the original essence of the data, and (ii) the temporal dimension is fundamental for understanding the sequence of movement.

The scope of this work is to propose a novel summarization method for big trajectory data with multiple aspects, aiming to provide a representative MAT (*RT*), i.e., from a set of filtered MATs (**T**) based on specific criteria[1], we intend to compute a single summarized MAT that refers to a representative MAT that provides a sequence of both spatial and temporal information while summarizing all aspects within MATs while reducing the data volume.

When computing RT, we are consequently faced with several challenges: (i) dealing with many heterogeneous aspects involved in **T**; (ii) reducing the information within the summarized trajectories while minimizing potential data loss; (iii) ensuring that RT encompasses essential details of the original data, leveraging the underlying spatial dimension and temporal sequence.

5 MAT-SGT: Multiple Aspect Trajectory Summarization based on a spatial Grid and Temporal sequence

This section introduces a novel method for computing representative MAT, named *MAT-SGT (Multiple Aspect Trajectory Summarization based on a spatial Grid and Temporal sequence)*. The development of this method is motivated by the existing literature gap in the field of MAT summarization. Recognizing that the computation of representative trajectories should align with the specific objectives and requirements of the intended use, our method focuses on computing representative MAT that captures the main behavior and characteristics of the input MATs, considering the spatiotemporal density and frequency of each aspect attribute value.

MAT data, with its three dimensions encompassing spatial, temporal, and semantic aspects, presents challenges in analyzing and extracting meaningful insights. To address this problem, our method analyzes the distribution of MAT points over time and space, enabling the identification of information values that best represent the main behavior exhibited in the input MATs. By leveraging spatiotemporal analysis techniques, we can capture patterns and trends in movement, providing valuable insights into the overall trajectory data with a focus on the spatiotemporal sequence.

To maintain representative MAT generated by *MAT-SGT*, we rely on a conceptual data model shown in Figure 2. This conceptual model provides a standardized representation of the input data and keeps the representative points as well as their mappings to the input points.

The conceptual model encompasses all dimensions of a MAT point. The spatial information is captured through the x and y coordinates. The temporal aspect can be represented either as a single timestamp or as a time interval denoting the start and end times. The semantic dimension is organized as a set of attributes associated with its corresponding value. These attributes can be categorical or numerical, providing insights into different characteristics or properties of the MAT point.

We also model *representative points*. It can encompass all the attributes of MAT points as a specialized class. A sequence of these representative points forms the final representative MAT, the RT. To compute RT, we summarize the information into *representative MAT points (p_r)*. Each one is derived by considering multiple input MAT points, and a relationship between the p_r and its corresponding MAT points is established and maintained to ensure accurate representation. Since p_r is a specialized point, it has the capability to hold all the attributes associated with the MAT points.

5.1 *MAT-SGT* Architecture

Figure 3 provides an overview of the *MAT-SGT* method, which comprises two main components: *(i) Data Segmentation* and *(ii) p_r Computation*. The first component aims to identify underlying data patterns based on data density (spatiotemporal), while the second focuses on summarizing the data by analyzing its frequency.

The method takes a set of filtered MATs (**T**) based on specific criteria (step 1). These criteria are out of the scope of this paper, but examples could encompass operations like clustering or straightforward filtering. For instance, in the context of this paper, these criteria might involve tasks such as given MATs generated by check-ins of different individuals to discern their patterns during specific time periods. An example of a simple filter could be restraining the dataset to contain only the trajectories of a particular individual during these defined time intervals. Then, the input MAT points are segmented into a cell grid (step 2) to identify relevant cells. For each relevant cell, representative points p_r's are computed (steps 4 to 6) that summarize all dimensions and capture essential input data characteristics.

During the *MAT-SGT* process, computed p_r are ordered by temporal dimension (step 7). This produces the RT output data (step 8). The best RT is selected in step 9. The best RT is determined by its similarity, coverage, and superiority over others in two new computations. To clarify the concepts of cell, relevant cell, and temporal interval, consider the example of analyzing delivery vehicle movements in a city. The spatial area is divided into a grid of cells, each representing a specific spatial region, such as a 500-meter by 500-meter section of the city where trajectory points are located. If multiple vehicles pass through this area, their trajectories are recorded in this cell. A relevant cell is one that exceeds a predefined threshold of activity—such as 50 or more vehicle visits in a day—making it significant for further analysis. Each trajectory point within a relevant cell is associated with time information. A Significant Temporal Interval (STI) highlights key periods of heightened activity within these temporal intervals. STIs allow us to focus on the most critical time frames of activity, ignoring less relevant periods. This example demonstrates how our approach efficiently captures and summarizes both spatial and temporal patterns, using rel-

[1] These criteria are out of the scope of this paper.

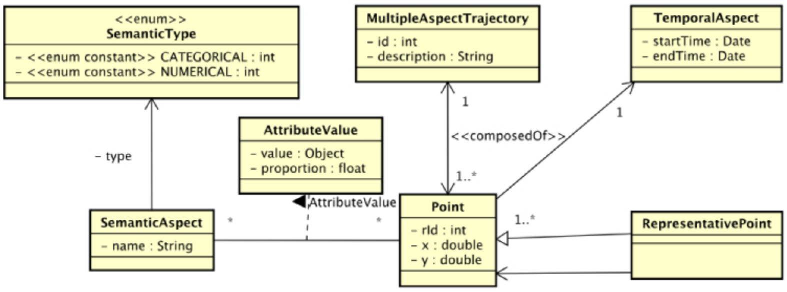


Figure 2. The conceptual model for *MAT-SGT*

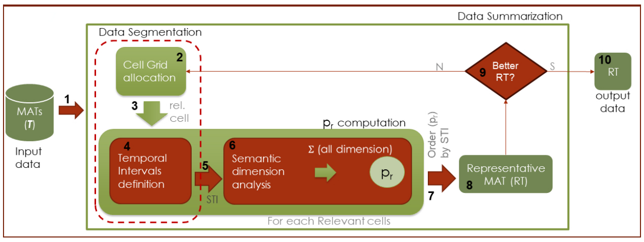


Figure 3. Overview of the *MAT-SGT* method.

evant cells and STIs to provide deeper insights into vehicle movement behavior.

Section 5.6 provides a detailed explanation of the selection process. *MAT-SGT* offers a comprehensive representation of behaviors and characteristics of input MATs, considering spatial and temporal density and frequency of attribute values. The following section details the *MAT-SGT* process.

5.2 Algorithm

The *MAT-SGT* algorithm is designed to consider a set of input parameters, outlined in Table 2. Analyst-defined parameters, τ_{rc} and τ_{rv}, contribute to the flexibility of the algorithm. In particular, τ_{rc} and τ_{rv} are optional parameters, allowing the analyst to tailor the method to specific needs. In contrast to our previous method (*MAT-SG*), which considers a constant value z for spatial grid cell size calculation, *MAT-SGT* introduces an automated procedure. This procedure iteratively analyzes representative trajectories for different values of z and selects the optimal value that yields the best RT. A detailed explanation of this iterative process follows.

Table 2. Parameters of *MAT-SGT*

Parameter	Explanation	Default
T	Set of previously filtered input MATs	-
τ_{rc}	Minimum proportion of all input MAT points $\|T.points\|$, deciding if a cell is considered a relevant cell to compute p_r	$rc = 2$
τ_{rv}	A rate of representativeness value for ranking values*	10%

* Ranking values are computed by data frequency, specifically only for the temporal dimension and categorical values of the semantic dimension.

The *MAT-SGT* algorithm (Algorithm 1) computes an RT by identifying the optimal spatial segmentation. It begins by computing the minimum spatial threshold (τ_s, given by the $computeMinSpatialThreshold()$ in line 2) to measure the dispersion between input points. It then determines the initial z value by calculating the distance between the grid origin (0,0) and the point furthest away from it ($computeMaxZValue()$ in line 5). Using this initial z value, a grid with a single cell containing all MAT points is generated (lines 9 and 10). It iteratively reduces the z value to compute a better RT (lines 8 to 26).

The algorithm aims to find the optimal segmentation for a better RT. In this step, a significant difference from the *MAT-SG* method becomes evident. In *MAT-SGT*, each iteration involves spatiotemporal data segmentation based on the current z value, providing spatial allocation (referred to as the *Cell Grid allocation* step). Representative points are then calculated by analyzing the temporal intervals for each group of points. This approach contrasts with *MAT-SG*, which relies solely on the segmentation of spatial data. Moreover, in *MAT-SGT*, the temporal sequence of representative points generates the RT. This distinction underscores the improved capability of *MAT-SGT* to capture both spatial and temporal nuances in the trajectory data.

As previously mentioned, *MAT-SGT* achieves MAT summarization through two internal components: *(i) data segmentation*; and *(ii)* p_r *computation*. The resulting RT quality is compared to the previous one ($betterRT$), and if an improvement of at least 10% is observed, $betterRT$ is updated (lines 17 to 21). The algorithm stops and returns the best RT if no improvements are detected in two iterations. The subsequent sections provide a detailed exploration of each component of the *MAT-SGT* method.

Algorithm 1: *MAT-SGT*

```
   input  : $\mathbf{T}$, $\tau_{rc}$, $\tau_{rv}$
   output : $RT$                                   /* representative trajectory */
1  $rc \leftarrow |\mathbf{T}.points| \times \tau_{rc}$;
2  $\tau_s \leftarrow$ computeMinSpatialThreshold();
3  $rt, betterRT \leftarrow \emptyset$
4  $betterRTmeasure, count \leftarrow 0$;
5  $z \leftarrow$ computeMaxZValue();
6  $w_{sim} \leftarrow 0.5$;
7  $w_{cover} \leftarrow (1 - w_{sim})$;
8  while $z > 1$ do
       // component (i) - Fig. 3 (steps 2 and 3)
9      $cellSize \leftarrow$ computeCellSize($\tau_s, z$);
10     $relCells \leftarrow$ cellGridAllocation($rc, cellSize$);
       // components (i) and (ii) - Fig. 3 (step 4 and 5)
11     $setGroupPoints \leftarrow$ STIdefinition($relCells, \tau_{rv}$);
       // component (ii) - Fig. 3 (step 6)
12     foreach $eachGroupPoint \in setGroupPoints$ do
13         $p_r \leftarrow$ computeRepPoint($eachGroupPoint, \tau_{rv}$);
14         $rt \leftarrow rt \cup p_r$
15     $rt.sort()$; // order by STI - Fig. 3 (step 7)
       // analysis of better $RT$ - Fig. 3 (step 9)
16     $rtMeasure \leftarrow RMMAT(rt, \mathbf{T}, w_{sim}, w_{cover})$;
17     if $(rtMeasure \times 1.1) \geq betterRTmeasure$ then
18         $betterRTmeasure \leftarrow rtMeasure$;
19         $betterRT \leftarrow rt$;
20         $rt \leftarrow \emptyset$
21         $count \leftarrow 0$;
22     else
23         $count++$;
24     if $count > 1$ then
25         break;
26     $z \leftarrow z \times 0.85$;
27 return $betterRT$;
```

5.3 Method Descriptions

- **computeCellSize(τ_s,z) - line 9:** This method calculates the size of the spatial cells based on the input parameters. Here, τ_s refers to the spatial dispersion threshold (τ_s), which defines the maximum allowable spatial distance between any two points within the same cell. This threshold represents the diagonal length of each cell, ensuring that all points within the cell remain within the specified spatial range. The method outputs the dimensions of each grid cell as a numerical value. x
- **cellGridAllocation(rc, cellSize) - line 10**: This method allocates each trajectory point to a specific cell in the spatial grid. Points are assigned to cells in the spatial grid based on their coordinates, creating a structured representation with trajectory points distributed across relevant cells.
- **STIdefinition(relCells, τ_{rv}) - line 11:** This method identifies *STIs* for each relevant cell by analyzing the temporal distribution of the points and ranking intervals based on frequency and occurrence. This process helps detect important temporal patterns in the trajectory data for temporal summarization.
- **computeRepPoint(eachGroupPoint, τ_{rv}) - line 13:** Based on the identified STIs, this method computes a *representative point* for each group of points within a relevant cell by aggregating the points within the same temporal group to derive a single representative point.

5.4 Data Segmentation Component

The data segmentation component of *MAT-SGT* operates in two crucial steps: *(i) Cell Grid Allocation* and *(ii) Temporal Intervals Definition*.

In the first step, the algorithm computes the cell size based on the given value of z and τ_s. This cell size determines the granularity of the spatial segmentation. The input MAT points are then allocated into the respective cells of the spatial grid. The algorithm identifies relevant cells containing at least rc points, ensuring sufficient data for meaningful representation and insights.

The second step involves the analysis of relevant cells to compute STIs for both data segmentation and the computation of representative points. For data segmentation, the STI rank is determined for each relevant cell. This process entails analyzing all temporal intervals within the cell and evaluating their tendency based on a frequency rate threshold of τ_{rv}, determining which intervals can be considered representative. The resulting STI rank encapsulates the temporal patterns present in the input MATs for each relevant cell.

Subsequently, the algorithm groups MAT points based on each $sti \in STI$ of its corresponding relevant cell (Algorithm 1, line 11). This grouping facilitates the identification and extraction of meaningful points with similar temporal characteristics, contributing to a comprehensive spatiotemporal representation of the trajectory data.

5.5 p_r Computation Component

The second component of *MAT-SGT* focuses on summarizing the groups of points obtained from the first component. This involves the computation of a representative point (p_r) for each group by summarizing the spatial, temporal, and semantic dimensions. These p_r's are sorted into a temporal sequence, ultimately constituting the RT.

The algorithm computes the centroid of the points within each group to determine the spatial dimension. In the temporal dimension, we use the sti technique we explained earlier. Different strategies are applied when dealing with semantic dimensions, which may encompass both *categorical* and *numerical* aspects.

For numerical aspects (e.g. temperature or air humidity), the algorithm computes the median value. In the case of categorical aspects (e.g. transportation means or weather conditions), a ranking of representative mode values is computed. *MAT-SGT* uses a predefined threshold (τ_{rv}) to determine which rank values are representative. After identifying these values, they are normalized to collectively sum to 100%, effectively representing the distribution of these values within the group, ensuring that the relative importance of the remaining values is properly reflected. This normalization ensures that the relative importance of the remaining categories is accurately reflected. Without this step, the remaining percentages would no longer represent the full dataset, leading to a misinterpretation of their significance. Normalizing the values provides a clearer and more meaningful summary, making it easier to interpret the overall results.

For the sake of understanding, consider a group of five data points with *POI* information: two points labeled *"restaurant"*, two points labeled *"university"*, and one point labeled *"library"*. Applying *MAT-SGT*, the initial mode values are *"restaurant"* and *"university"*, each representing 40% of the data, while *"library"* accounts for 20%. With a representative value threshold of $\tau_{rv} = 25\%$, the *"library"* value is excluded. The proportions of *"restaurant"* and *"university"* are updated, with each now representing 50% of the representative values. This reorganization ensures an accurate representation of values within the group, summarizing categorical data. The p_r computation step combines centroids, sti, and representative values for numerical and categorical aspects, contributing to determining the RT.

5.6 Computation of the Better Representative Trajectory

To analyze and compute the better *RT* (according to Figure 3, step 9), *MAT-SGT* employs a representativeness measure called RMMAT [Machado *et al.*, 2023b], the state-of-the-art for representativeness measure for MAT summarization. RMMAT reflects the overall coverage of both MAT points and the information in the RT. This measure is computed in line 16 in Algorithm 1. This analysis sets w_{sim} and w_{cover} to equal values. Combining the similarity measure and coverage proportion, RMMAT aims to identify the RT that achieves the maximum coverage of both MAT points and their contained information. The similarity measure is computed using MUITAS [Petry *et al.*, 2019], recognized as the state-of-the-art measure for MATs.

The *MAT-SGT* method prioritizes spatiotemporal segmentation. This means that if all points within the same cell are semantically different, the algorithm analyzes the temporal density of the points. It computes at least one representative point considering spatial and temporal dimensions. This approach emphasizes the representativeness of a specific location at a particular time in the input MATs. By incorporating temporal density analysis, the method captures the significance of an area at a particular moment, considering the dynamic nature of the data.

5.7 Output data

The output data (Figure 3, step 10) comprises a representative MAT denoted as RT, presented in the form of a CSV file. Two main components determine the structure of the CSV file: (i) the configuration settings for the RT computation and (ii) the information associated with each representative MAT point.

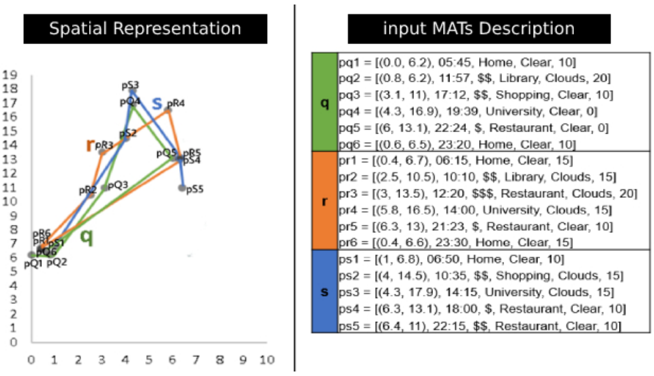
The configuration settings include the following data: $CellSize$, τ_{rc}, τ_{rv}, $|cell|$, $minPointRC$, $|RT|$, and $|coverPoints|$. Each data serves a specific purpose:

- $CellSize$ refers to the final cell size of the spatial grid;
- $|cell|$ refers to the number of cells that were computed in the model;
- $minPointRC$ refers to the minimum number of points that are needed in each cell to be considered relevant in the RT computation;
- $|RT|$ refers to the size of RT, which is the number of p_r's;
- $|coverPoints|$ refers to the number of input MAT points that the RT cover, as determined by the mapping information.

The second element in the output file contains information about each representative MAT point (p_r). This information is structured as: "lat_lon, $time$, $\#Semantic\_Aspects\#$, $mapping$". The "lat_lon" refers to the spatial dimensions of the point, comprising latitude and longitude. The "$time$" refers to the temporal aspects of the point, which can be either an interval or a single occurrence. The "$\#Semantic\_Aspects\#$" illustrated all the semantic aspects present in the input MATs, encompassing categorical or numerical types. Categorical types are expressed as a normalized rank of information, such as weather conditions represented as "{CLOUDS: 0.5; CLEAR: 0.4; RAIN: 0.1}". Numerical types are represented by their median value. Finally, the "$mapping$" refers to the input MAT points that make up the referent p_r. For instance, in "127: 3; 127: 9; 129: 43; 134: 92; 137: 110; 137: 118; 138: 139," indicates that the current p_r is composed of points with ID #3 and #9 from the trajectory ID #127, along with other points.

5.8 Running Example

Let's give an example to illustrate how *MAT-SGT* works. We consider a set of input MATs $\mathbf{T} = \langle q, r, s \rangle$, where $q = \langle p_{q_1}, p_{q_2}, ..., p_{q_n} \rangle$, $r = \langle p_{r_1}, p_{r_2}, ..., p_{r_m} \rangle$ and $s = \langle p_{s_1}, p_{s_2}, ..., p_{s_t} \rangle$. Figure 4 presents the trajectories and their related aspects. To enhance clarity, each MAT point is described in the figure (on the right side) by the following attributes: spatial information, occurrence time, *price* spent at PoIs (when having value), visited *PoIs*, the *weather conditions*, and the *rain precipitation*, where the spatial information refers to the spatial coordinates (latitude and longitude). Additionally, to address the issue of overlapping lines in the figure, we have adjusted the visualization to improve the distinction between trajectories. This includes using different colors for each trajectory, making it easier to follow the paths without confusion.



q	pq1 = [(0.0, 6.2), 05:45, Home, Clear, 10]
	pq2 = [(0.8, 6.2), 11:57, $$, Library, Clouds, 20]
	pq3 = [(3.1, 11), 17:12, $$, Shopping, Clear, 10]
	pq4 = [(4.3, 16.9), 19:39, University, Clear, 0]
	pq5 = [(6, 13.1), 22:24, $, Restaurant, Clear, 0]
	pq6 = [(0.6, 6.5), 23:20, Home, Clear, 10]

r	pr1 = [(0.4, 6.7), 06:15, Home, Clear, 15]
	pr2 = [(2.5, 10.5), 10:10, $$, Library, Clouds, 15]
	pr3 = [(3, 13.5), 12:20, $$$, Restaurant, Clouds, 20]
	pr4 = [(5.8, 16.5), 14:00, University, Clouds, 15]
	pr5 = [(6.3, 13), 21:23, $, Restaurant, Clear, 10]
	pr6 = [(0.4, 6.6), 23:30, Home, Clear, 15]

s	ps1 = [(1, 6.8), 06:50, Home, Clear, 10]
	ps2 = [(4, 14.5), 10:35, $$, Shopping, Clouds, 15]
	ps3 = [(4.3, 17.9), 14:15, University, Clouds, 15]
	ps4 = [(6.3, 13.1), 18:00, $, Restaurant, Clear, 10]
	ps5 = [(6.4, 11), 22:15, $$, Restaurant, Clear, 10]

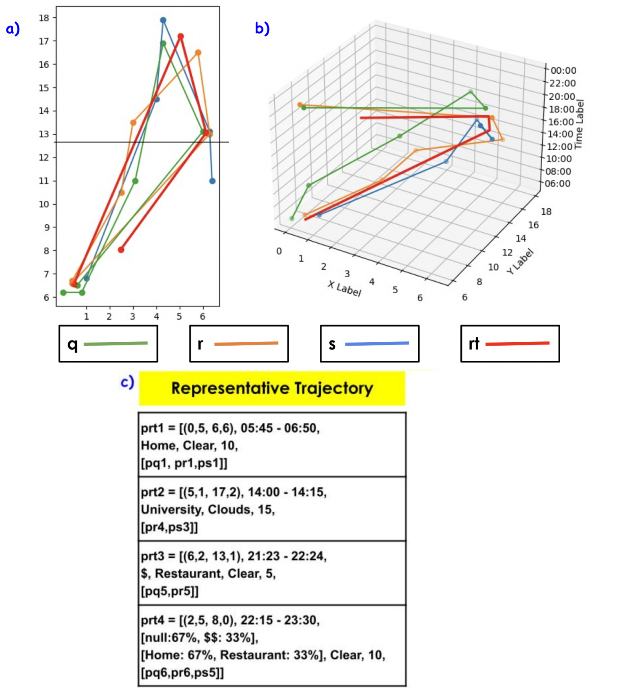
Figure 4. Sample data illustrating trajectories, each delimited in different colors: trajectory q (green), r (orange), and s (blue). On the left side, the trajectories are displayed in a spatial plan, while on the right side, each trajectory point is described. The format for each point includes: [(latitude, longitude), time, price of the PoI, PoI information, weather condition, rain precipitation].

We consider $\tau_{rc} = 25\%$ and $\tau_{rv} = 25\%$ as input values. As $|\mathbf{T}.points| = 17$, a relevant cell must contain more than 4 points. Figure 5 shows the resulting $rt = \langle p_{rt_1}, p_{rt_2}, p_{rt_3}, p_{rt_4} \rangle$ (red line) in different perspectives. Figure 5 (a) shows the spatial distribution of the representative trajectory computed from $\mathbf{T}$. The input MATs are segmented into a grid of cells, and the red line indicates the corresponding RT. Figure 5 (b) illustrates a spatiotemporal perspective displaying the evolution of the input MATs and the computed RT, providing insights into how they unfold over time. Detailed output is illustrated in Figure 5 (c), providing additional information and insights about the RT. As stated before, data summarization occurs within cells that contain more than one point.

Figure 5 (c) detail each representative point in representative MAT, here each p_{rt} is described by: spatial information, time intervals (sti), *price* spent at PoIs (when having value), visited *PoIs*, the *weather conditions*, the *rain precipitation*, and mapping information. The mapping information outlines the derived MAT points from the input dataset. Specifically, p_{rt_1} is derived from p_{q1}, p_{r1} and p_{s1}.

Additionally, p_{rt_1} and p_{rt_4} are derived from the first cell (as shown in the more down cell of Figure 5 (a)), while p_{rt_2} and p_{rt_3} are derived from the second cell.

In the first cell, we identify two important time intervals (sti) from the input MATs during the *Temporal Intervals definition* step. The first sti covers the time interval between 05:45 and 05:50, indicating a period when significant activity was recorded. The second sti spans from 22:15 to 23:30, representing another critical temporal interval. These sti's contain critical MAT points that contribute to the computa-



prt1 = [(0,5, 6,6), 05:45 - 06:50,
Home, Clear, 10,
[pq1, pr1,ps1]]

prt2 = [(5,1, 17,2), 14:00 - 14:15,
University, Clouds, 15,
[pr4,ps3]]

prt3 = [(6,2, 13,1), 21:23 - 22:24,
$, Restaurant, Clear, 5,
[pq5,pr5]]

prt4 = [(2,5, 8,0), 22:15 - 23:30,
[null:67%, $$: 33%],
[Home: 67%, Restaurant: 33%], Clear, 10,
[pq6,pr6,ps5]]

Figure 5. Resulting in representative trajectory (RT) visualization in different perspectives: (a) Spatial; (b) Spatiotemporal; and (c) RT description.

tion of RT, with p_{rt_1} representing the referent MAT point for the first segment and p_{rt_4} for the second segment in the same cell, both derived from the identified time intervals.

In short, the *MAT-SGT* method aims to compute an RT that effectively captures the main behavior and characteristics of the input MATs. It achieves this by considering the spatiotemporal density and frequency of each aspect attribute value. By analyzing the distribution of MAT points over time and space, the method identifies and prioritizes the significant segments and aspects, leading to an RT that represents the key features of the input MATs.

6 Experimental Evaluation

This section presents an experimental evaluation of the proposed summarization method *MAT-SG* and its comparison with the MAT state-of-the-art *MAT-SG*, shedding light on their utility and representativeness. *MAT-SG* was chosen in this comparison because both methods consider the individual analysis of semantic data as categorical or numerical, and both methods include mappings between input MATs and the representative MAT, allowing the use of the coverage data in the analysis. All experiments were implemented in Java and conducted on a Dell Inspiron laptop with an Intel Core i5 processor and 16 GB memory. In the following sections, we describe the datasets (Section 6.1), the experimental setup (Section 6.2), and the results (Section 6.3).

6.1 Datasets

We evaluate the effectiveness of our method using three datasets containing MATs, Foursquare, Gowalla, and Brightkite. These datasets, widely employed in other works [Petry *et al.*, 2019; da SILVA *et al.*, 2019; Tortelli Portela *et al.*, 2022], contribute to the robustness of our evaluation.

Table 3. Summary of the used datasets

Dataset	Description	Aspects
Foursquare	Traj Size: ~ 22 # of Traj.: 3079 # of Points: 66962 # filtered data groups: 193 Filter Criteria: User	Lat, Lon, Time, Weather Conditions, PoI - Category, Price, and Rating
Gowalla	Traj Size: ~ 18 # of Traj.: 5329 # of Points: 98158 # filtered data groups: 300 Filter Criteria: User	Lat, Lon, Time, PoI, Weekday
Brightkite	Traj Size: ~ 16 # of Traj.: 7911 # of Points: 130494 # filtered data groups: 300 Filter Criteria: User	Lat, Lon, Time, PoI, Weekday

The diversity in these datasets ensures a comprehensive evaluation, considering multiple dimensions and aspects of trajectory data.

The *Foursquare NYC dataset* is a well-established trajectory dataset encompassing check-in data in New York City, spanning from April 2012 to February 2013. This dataset not only includes *spatial* and *temporal* information but also incorporates some semantic aspects such as *weekday*, *weather conditions*, and aspects like *category*, *price*, and *rating* of Points of Interest (POIs). With a total of 3079 trajectories from 193 users, the dataset presents a rich set of approximately 22 check-ins per trajectory, with an average of approximately 16 trajectories per user.

The *Gowalla Location-Based Social Network* is a dataset collected worldwide between February 2009 and October 2010. For our analysis, we used 300 random users and limited the trajectory sizes between 10 and 50 check-ins, resulting in 5329 trajectories. This dataset provides information about *anonymized users*, *POIs*, *spatial*, and *temporal* details, along with enriched semantic information about *weekdays*.

The *Brightkite dataset*, sourced from the Brightkite social media platform and collected between April 2008 and October 2010 Cho *et al.* [2011], includes a randomly selected subset of 300 users. The dataset comprises a total of 7911 trajectories, each with a consistent range of 10 to 50 points. It comprises the exact dimensions of the Gowalla dataset, including the enriched semantic information of the weekday.

Table 3 presents the characteristics of each dataset, with the average trajectory size, the number of trajectories, points, filtered data groups, and the filter criteria used for each dataset.

6.2 Experimental Setup

Our experimental evaluation takes a systematic approach to assess the utility of the representative MAT (RT) by using two distinct strategies: (i) the RMMAT measure and (ii) the Average Recall (AR) metric.

6.2.1 RMMAT Strategy

In our first strategy for evaluating the RT, we leverage recent research and employ the state-of-the-art representativeness measure for MAT summarization, specifically the RMMAT measure [Machado *et al.*, 2023b]. This measure focuses on evaluating the *RT* across the entire dataset, aiming to comprehensively evaluate its quality in terms of both similarity and coverage of information.

The RMMAT measure involves the computation of *RT* for each filtered trajectory group (**T**) within the dataset (**D**). The dataset is effectively segmented into multiple groups ($T \in \mathbf{T} \in \mathbf{D}$), allowing for a detailed evaluation of representativeness within different subsets.

The RMMAT measure produces a value within the range of 0 to 1, where a value closer to 1 indicates that the *RT* effectively represents the dataset, while a value closer to 0 suggests that the *RT* contains less or no information from the dataset. To ensure a balanced consideration of both similarity and covered information, we adopt a strategy with equal weights, setting $\omega_{sim} = \omega_{cover} = \frac{1}{2}$. This approach allows for a more comprehensive evaluation of representativeness.

6.2.2 AR Metric Strategy

In addition to the RMMAT evaluation, we employ a second strategy using the Average Recall (AR) metric. Inspired by the work of similarity measure MUITAS [Petry *et al.*, 2019], where our adoption of the AR metric aligns with their methodology. However, our focus remains distinct — we aim to evaluate the utility of *RT* within the context of the input dataset, quantifying the quality of our summarization and representative data computation.

The AR metric measures recall based on the similarity between the *RT* computed by *MAT-SGT* and other trajectories in the dataset. The entire dataset (**D**) is divided into multiple groups ($T \in \mathbf{T} \in \mathbf{D}$), and the *RT* is computed for each group. It is assumed that trajectories within the same group exhibit similarity, and we aim for high similarity values between the *RT* and trajectories within the same group. The evaluation process unfolds systematically. The *RT* is computed for each group, and a similarity search is conducted over the dataset. Trajectories are ordered by similarity, and recall is calculated. The assessment relies on the ideal scenario where the top k most similar trajectories align with the trajectories of the same group, represented as $k = |T_{group}|$. This indicates that all trajectories within the same group are considered the most similar to the *RT* of that group, effectively gauging the *RT*'s ability to rank trajectories accurately within the same ground truth group.

The AR can be formally defined as follows:

$$AR = \frac{|\text{Relevant Trajectories} \cap \text{Retrieved Trajectories}|}{|\text{Relevant Trajectories}|} \quad (1)$$

In this formula, the numerator represents the count of relevant trajectories ($|T_{group}|$) that are successfully retrieved (k), while the denominator indicates the total number of relevant trajectories within the cluster. A higher AR value signifies better performance in ranking the *RT* in alignment with the actual trajectories of the same group.

6.2.3 Experimental Evaluation

We performed experiments by executing *MAT-SG* and *MAT-SGT* in each ground truth (i.e., each user, as criteria definition to filter trajectories into groups). The method was repeated on each user with a different setting of the parameters τ_{rv} and τ_{rc} with values varying from 5% to 25%, resulting in 25 runs for each user. This parameter variation allows for evaluating the sensitivity and robustness of the methods.

To compute the similarity measure between trajectories, we rely on MUITAS [Petry *et al.*, 2019], the state-of-the-art w.r.t. MAT similarity measure. Proximity functions assess spatial, temporal, and semantic matching between $T \in \mathbf{T}$ and *RT*, considering the distinct structure of *RT*. We use the Euclidean distance measure for spatial matching, considering $2 \times$ cellSize as the threshold. For temporal matching, we use the timestamp value of T falling within the interval of *RT*. For semantic matching, we evaluate attribute values for *numeric* and *categorical* types. A numerical match considers a threshold of 10% of the *RT* value, while a categorical match considers if the value of T is within the range of values of *RT*. We set $w = 1/3$ for each dimension of MUITAS to balance all of them.

6.3 Results

6.3.1 RMMAT Strategy

We analyze the representativeness measure of each *RT* computed by both methods *MAT-SG* and *MAT-SGT* for different specified parameter configurations. The parameters τ_{rv} and τ_{rc} are utilized to represent the x-axis (each row in the tables) and y-axis (each column), respectively. Higher values indicate better representativeness, and we highlight them in bold. Conversely, the lowest values are underlined. We are comparing the performance of two models: *MAT-SG* and *MAT-SGT*.

Foursquare-NYC dataset Tables 4 and 5 present the average RMMAT results for *RT* computations with different parameter configurations. The highest representativeness measures were obtained with *MAT-SG* (0.692) and *MAT-SGT* (0.627), both with τ_{rv} and τ_{rc} set to 0.05. Conversely, the lowest values were recorded for both methods (0.201 for *MAT-SG* and 0.207 for *MAT-SGT*), with τ_{rv} and τ_{rc} both set to 0.25. The average of RMMAT was found to be superior under the best configuration in *MAT-SG*.

Table 4. Average of RMMAT of user trajectories in Foursquare dataset by *MAT-SG*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.692**	0.573	0.466	0.363	0.249
0.1	0.663	0.543	0.438	0.339	0.232
0.15	0.637	0.515	0.412	0.318	0.217
0.2	0.616	0.494	0.393	0.303	0.207
0.25	0.600	0.481	0.383	0.295	0.201

Gowalla dataset Tables 6 and 7 insert show results for the Gowalla dataset. The highest representativeness measures were obtained with *MAT-SG* (0.693) and *MAT-SGT* (0.624),

Table 5. Average of RMMAT of user trajectories in Foursquare dataset by *MAT-SGT*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.627**	0.553	0.501	0.476	0.480
0.10	0.561	0.479	0.423	0.394	0.380
0.15	0.498	0.403	0.334	0.296	0.273
0.20	0.443	0.346	0.277	0.248	0.238
0.25	0.402	0.305	0.243	0.223	0.207

both with τ_{rv} and τ_{rc} set to 0.05. Conversely, the lowest values were recorded for both methods (0.238 for *MAT-SG* and 0.225 for *MAT-SGT*), with τ_{rv} and τ_{rc} both set to 0.25. The average of RMMAT was found to be superior under the best configuration in *MAT-SG*.

Table 6. Average of RMMAT of user trajectories in Gowalla dataset by *MAT-SG*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.693**	0.576	0.484	0.403	0.322
0.1	0.660	0.539	0.449	0.373	0.298
0.15	0.627	0.505	0.418	0.345	0.275
0.2	0.592	0.468	0.385	0.316	0.251
0.25	0.566	0.444	0.364	0.300	0.238

Table 7. Average of RMMAT of user trajectories in Gowalla dataset by *MAT-SGT*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.624**	0.555	0.524	0.505	0.499
0.1	0.558	0.474	0.438	0.407	0.391
0.15	0.487	0.391	0.351	0.319	0.310
0.2	0.424	0.320	0.283	0.256	0.248
0.25	0.377	0.283	0.252	0.228	0.225

Brightkite dataset Tables 8 and 9 display the average RMMAT results for the Brightkite dataset. The highest representativeness measures were obtained with *MAT-SG* (0.875) and *MAT-SGT* (0.738), both with τ_{rv} and τ_{rc} set to 0.05. Conversely, the lowest values were obtained by *MAT-SG* (0.551) with τ_{rv} and τ_{rc} both set to 0.25, and by *MAT-SGT* (0.298) with $\tau_{rv} = 0.25$ and $\tau_{rc} = 0.15$.

Table 8. Average of RMMAT of user trajectories in Brightkite dataset by *MAT-SG*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.875**	0.834	0.797	0.771	0.723
0.1	0.820	0.774	0.735	0.709	0.662
0.15	0.771	0.722	0.681	0.656	0.609
0.2	0.726	0.676	0.636	0.612	0.566
0.25	0.705	0.656	0.618	0.594	0.551

6.3.2 AR Metric Strategy

This section evaluates the Average Recall (AR) metric for ranking user trajectories within the same group based on a specified parameter configuration. The parameters τ_{rv} and τ_{rc} are employed, representing the x-axis (each row in the tables) and y-axis (each column), respectively. Higher values indicate better exactness, highlighted in bold, while the lowest values are underlined. We compare the performance of two models: *MAT-SG* and *MAT-SGT*.

Table 9. Average of RMMAT of user trajectories in Brightkite dataset by *MAT-SGT*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.738**	0.716	0.693	0.675	0.648
0.1	0.564	0.518	0.495	0.481	0.457
0.15	0.447	0.399	0.385	0.390	0.393
0.2	0.366	0.341	0.334	0.349	0.354
0.25	0.329	0.299	0.298	0.317	0.320

Foursquare-NYC dataset Tables 10 and 11 display the results for ranking user trajectories using AR. For *MAT-SG*, the highest value of 0.785 occurs with τ_{rv} and τ_{rc} both set to 0.05, while the lowest value (0.450) is obtained with τ_{rv} and τ_{rc} both set to 0.25. *MAT-SGT* achieves the highest value of 0.848 with both parameters set to 0.05, while the lowest value (0.372) is obtained with $\tau_{rv} = 0.25$ and $\tau_{rc} = 0.15$ highlighting its effectiveness under the best parameter configuration.

Table 10. AR of ranking user trajectories in Foursquare dataset by *MAT-SG*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.785**	0.643	0.546	0.498	0.472
0.1	0.770	0.627	0.534	0.483	0.475
0.15	0.743	0.600	0.521	0.471	0.460
0.2	0.742	0.609	0.526	0.478	0.456
0.25	0.734	0.599	0.524	0.473	0.450

Table 11. AR of ranking user trajectories in Foursquare dataset by *MAT-SGT*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.848**	0.755	0.686	0.641	0.634
0.1	0.809	0.680	0.592	0.534	0.517
0.15	0.731	0.573	0.475	0.431	0.420
0.2	0.656	0.490	0.410	0.400	0.394
0.25	0.586	0.432	0.372	0.377	0.388

Gowalla dataset Tables 12 and 13 provide the corresponding results for the Gowalla dataset. For *MAT-SG*, the highest value (0.871) is achieved with both τ_{rc} and τ_{rv} set to 0.05, while the lowest value (0.546) is identified with both parameters set to 0.25. On the other hand, *MAT-SGT* achieves the highest AR (0.888) with both τ_{rc} and τ_{rv} set to 0.05, and the lowest value (0.509) is obtained with $\tau_{rc} = 0.25$ and $\tau_{rv} = 0.2$. Thus, in the Gowalla dataset, *MAT-SGT* demonstrates superior performance under the best parameter configuration.

Brightkite dataset Results for the Brightkite dataset are shown in Tables 14 and 15. *MAT-SG* achieves the highest

Table 12. AR of ranking user trajectories in Gowalla dataset by *MAT-SG*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.871**	0.771	0.729	0.702	0.692
0.1	0.838	0.724	0.669	0.639	0.633
0.15	0.807	0.682	0.620	0.589	0.566
0.2	0.753	0.646	0.601	0.572	0.546
0.25	0.732	0.663	0.643	0.634	0.609

Table 13. AR of ranking user trajectories in Gowalla dataset by *MAT-SGT*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.888**	0.826	0.804	0.799	0.795
0.1	0.865	0.771	0.732	0.710	0.693
0.15	0.794	0.664	0.608	0.595	0.575
0.2	0.690	0.558	0.519	0.513	0.509
0.25	0.644	0.537	0.518	0.515	0.517

AR (0.928) with both τ_{rc} and τ_{rv} set to 0.05, while the lowest value (0.819) is identified with $\tau_{rv} = 0.15$ and $\tau_{rc} = 0.25$. *MAT-SGT* attains the highest AR (0.954) with both parameters set to 0.05, and the lowest value (0.621) is obtained with τ_{rv} = 0.25 and τ_{rc} = 0.05. *MAT-SGT* showcasing superior performance under the best parameter configuration.

Table 14. AR of ranking user trajectories in Brightkite dataset by *MAT-SG*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.928**	0.905	0.898	0.884	0.869
0.1	0.920	0.897	0.890	0.873	0.860
0.15	0.887	0.871	0.857	0.838	0.819
0.2	0.866	0.860	0.863	0.847	0.841
0.25	0.865	0.859	0.867	0.854	0.845

Table 15. AR of ranking user trajectories in Brightkite dataset *MAT-SGT*

τ_{rc} τ_{rv}	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**
0.05	**0.954**	0.935	0.927	0.915	0.903
0.1	0.881	0.866	0.863	0.855	0.843
0.15	0.756	0.736	0.750	0.759	0.783
0.2	0.658	0.677	0.696	0.716	0.754
0.25	0.621	0.628	0.663	0.698	0.744

6.4 Discussion

Our evaluation focused on the summarization methods of dual capabilities: (i) ensuring the representativeness of the computed representative data for each input dataset using the RMMAT Metric and (ii) effectively ranking filtered trajectories using the AR Metric.

First, regarding RMMAT analysis, we evaluated our computation methods for *RT* in various scenarios and obtained an overall RMMAT score by observing the best parameter configuration. The results are presented in Table 16. Overall, both methods (*MAT-SG* and *MAT-SGT*) exhibited high RMMAT scores when considering the best parameter configuration by each user, indicating the effectiveness of our methods in summarizing user trajectories. Additionally, in most cases, *MAT-SG* outperformed *MAT-SGT* regarding the representativeness value across input data. For example, on the Brightkite dataset, *MAT-SG* achieved an average RMMAT score of 0.875, outperforming *MAT-SGT* 0.738.

Table 16. The compiled results of RMMAT across all experimental evaluations

Dataset	**Method**	**Best By User**		**All Results**					**Complete**	**Incomplete**
		AR	**Median**	**AR**	**Median**	**SD**	**Max.**	**Min.**		
Forsquare	***MAT-SG***	0.691	0.720	0.425	0.470	0.267	0.96	0.000	4800	0 + (2 users)
	MAT-SGT	0.637	0.640	0.390	0.400	0.201	0.940	0.000	4581	219 + (1 users)
Gowalla	***MAT-SG***	0.693	0.710	0.435	0.480	0.270	0.970	0.000	7375	0 + (5 users)
	MAT-SGT	0.632	0.630	0.395	0.400	0.207	0.900	0.000	7044	331 + (5 users)
Brightkite	***MAT-SG***	0.874	0.890	0.699	0.720	0.166	0.990	0.000	3850	0 + (146 users)
	MAT-SGT	0.739	0.745	0.475	0.500	0.243	0.980	0.000	3162	688 + (146 users)

The analysis suggests that *MAT-SG* demonstrates superior RMMAT values, especially when considering a parameter configuration aligned with the data pattern. This indicates at *MAT-SG* being more effective in certain scenarios for capturing the representativeness of trajectories, leading to better similarity and covered information.

One hypothesis is that, regarding similarity, using MUITAS that does not consider the sequence in data may be positive in *MAT-SG*. At the same time, it may not be the best measure in *MAT-SGT* since the temporal sequence is not considered in this measure. Currently, no similarity measures are available to compare data sequences for MATs. Regarding covered information, *MAT-SG* only considers the spatial dimension in segmentation, which means that more data points are summarized in each representative point. In contrast, since *MAT-SGT* considers two steps to segment data for spatial and temporal dimensions, the number of data points considered for computing the representative point is lower, providing a straightforward lower covered information.

Regarding AR analysis, in general, *MAT-SG* exhibits a linear AR result when ranking user trajectories for each τ_{rc} across a range of τ_{rv}. As τ_{rc} decreases, AR tends to decrease due to the algorithm's minimum requirement of MAT points in each cell for relevance. Conversely, *MAT-SGT* maintains a linear AR result for each τ_{rv} across a range of τ_{rc}, with decreasing AR as τ_{rv} decreases.

As the minimum requirement increases, it becomes more challenging to accurately rank user trajectories, leading to a decrease in the AR. When more MAT points are required to compute the representative MAT (*RT*), the algorithms have less power to rank the user's trajectories accurately. Additionally, when no cell is identified as relevant, the algorithms do not compute a p_r for the points in that cell.

The analysis of the results shows that the best values for τ_{rc} are around 0.05, with decreasing values of AR as τ_{rc} increase, suggesting the effectiveness of larger cell sizes in capturing group characteristics. Smaller cell sizes and stricter relevance criteria pose challenges for computing an *RT* that performs well across different scenarios.

Our *RT* computation methods were evaluated in various scenarios and achieved an overall AR score by observing the best parameter configuration. Results are presented in Table 17. Furthermore, the highest AR values achieved with the best parameter configurations indicate the superior performance of *MAT-SGT* in ranking user trajectories across datasets.

Table 17. The compiled results of AR across all experimental evaluations

Dataset	Method	Best By User		All Results					Complete	Incomplete
		AR	Median	AR	Median	SD	Max.	Min.		
Forsquare	*MAT-SG*	0.833	0.900	0.568	0.600	0.315	1.000	0.000	4800	0 + (2 users)
	MAT-SGT	0.886	0.930	0.560	0.600	0.324	1.000	0.000	4581	219 + (1 users)
Gowalla	*MAT-SG*	0.889	0.950	0.677	0.750	0.294	1.000	0.000	7375	0 + (5 users)
	MAT-SGT	0.909	0.960	0.672	0.730	0.295	1.000	0.000	7044	331 + (5 users)
Brightkite	*MAT-SG*	0.954	1.000	0.870	0.930	0.187	1.000	0.000	3850	0 + (146 users)
	MAT-SGT	0.966	1.000	0.797	0.900	0.252	1.000	0.000	3162	688 + (146 users)

We can observe that in some cases, in both analysis, there is insufficient density to determine a behavioral pattern (*Incomplete* column), where *MAT-SGT* has identified more incomplete *RT* across some parameter configurations. This situation arises due to the dual-step density segmentation employed by *MAT-SGT*. Therefore, more information is needed to analyze its representative data. Additionally, it is essential to consider different configurations because users exhibit different behavioral patterns.

These experimental evaluations provide a comprehensive and nuanced tool to understand both methods *MAT-SG* and *MAT-SGT* and represent filtered trajectories. Both methods demonstrate high effectiveness, with flexibility in adapting to individual group behavior patterns. This adaptability is particularly valuable for personalized services and targeted interventions. The choice between the two methods depends on the specific goal, where *MAT-SGT* excels when temporal information is critical to understanding the chronology of events or movements over time.

7 Conclusion

This paper introduced the *MAT-SGT* method for summarizing trajectories with multiple aspects and providing representative data. The effectiveness of computing an RT depends on its intended purpose. However, previous methods, such as the FSM-based approach Seep and Vahrenhold [2019] and MAT-SG Machado *et al.* [2022], had limitations in capturing temporal sequences. To address these limitations, *MAT-SGT* treats semantic types individually and identifies temporal sequences within movement patterns. It provides representative data and allows for identifying patterns and assessing data representativeness.

The AR metric evaluation highlights the effectiveness of *MAT-SGT* in capturing similarity between RT and other trajectories. Notably, it is important to mention that regarding the comparison of *MAT-SGT* and the previous works, there are some challenges, since in Seep and Vahrenhold [2019] presents unavailable source and insufficient information provided in the short article, it also lacked output data. Furthermore, it is significant to highlight the distinctive goals of *MAT-SG* and *MAT-SGT*. *MAT-SG* aims to identify representative spatial areas, while our proposed method *MAT-SGT* focuses on identifying representative data with both spatial and temporal dimensions. These relevant differences between both methods were highlighted during our analysis, where the *MAT-SGT* indicates the superior performance in ranking user trajectories across datasets, highlighting its potential when temporal information is critical to understanding the chronology of events or movements over time.

Our experiments provide insights into the performance of *MAT-SGT* and underscore the significance of parameter selection for optimal results. Parameter selection significantly impacts the quality and utility of RTs, emphasizing the need for careful tuning to achieve optimal results, as well the importance of considering the perspective of the analyst to achieve better RT. In this way, future work aims to refine the parameter selection process to enhance the method's performance in diverse datasets and real-world scenarios.

Additionally, we aim to enhance the management of temporal overlaps among intervals in different cells, as the current analysis does not adequately address this issue. This could lead to overlapping significant time intervals that may distort the representative trajectory. Our future research will explore advanced temporal clustering and spatial analysis techniques, enabling us to differentiate better and consolidate overlapping points.

Acknowledgements

This work has been partially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, SoBigData++ Project - by Transnational Access (TNA), as well as the European Union's Horizon 2020 research and innovation programme under GA N. 777695 (EU Project MASTER - Multiple ASpects TrajEctoRy management and analysis). The views and opinions expressed in this article are the authors' sole responsibility and do not necessarily reflect the views of the European Commission.

Funding

This research was partially funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, SoBigData++ Project - by Transnational Access (TNA), as well as the European Union's Horizon 2020 research and innovation programme under GA N. 777695 (EU Project MASTER - Multiple ASpects TrajEctoRy management and analysis).

Authors' Contributions

Vanessa Lago Machado is the main contributor and writer of this manuscript. Tarlis and Ronaldo contributed to defining the *MAT-SGT* measure, while Geomar contributed to the validation step. All authors have read and approved the final version of the manuscript.

Competing interests

The authors declare that they do not have any competing interests.

Availability of data and materials

The source code for *MAT-SGT* is available at `https://github.com/RepresentantativeMAT/MAT-SGT.git`. The datasets used during the current experimental evaluation are available in `https://github.com/bigdata-ufsc/datasets`.

References

Ahmed, M. (2019). Data summarization: a survey. *Knowledge and Information Systems*, 58(2):249–273.

Buchin, K., Buchin, M., Van Kreveld, M., Löffler, M., Silveira, R. I., Wenk, C., and Wiratma, L. (2013). Median trajectories. *Algorithmica*, 66(3):595–614.

Buchin, M., Kilgus, B., and Kölzsch, A. (2019). Group diagrams for representing trajectories. *International Journal of Geographical Information Science*, 34(12):2401–2433.

Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1082–1090. DOI: 10.1145/2020408.2020579.

da SILVA, C., PETRY, L., and BOGORNY, V. (2019). A survey and comparison of trajectory classification methods. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 788–793, Brazil. IEEE. DOI: 10.1109/BRACIS.2019.00141.

Ermakova, L., Cossu, J. V., and Mothe, J. (2019). A survey on evaluation of summarization methods. *Information processing & management*, 56(5):1794–1814.

Erwig, M., Schneider, M., Vazirgiannis, M., *et al.* (1999). Spatio-temporal data types: An approach to modeling and querying moving objects in databases. *GeoInformatica*, 3(3):269–296.

Etienne, L., Devogele, T., Buchin, M., and McArdle, G. (2016). Trajectory box plot: A new pattern to summarize movements. *Int. J. Geogr. Inf. Sci.*, 30(5):835–853.

Fiore, M., Katsikouli, P., Zavou, E., Cunche, M., Fessant, F., Le Hello, D., Aivodji, U., Olivier, B., Quertier, T., and Stanica, R. (2020). Privacy in trajectory micro-data publishing: a survey. *Transactions on Data Privacy*, 13:91–149.

Gao, C., Zhao, Y., Wu, R., Yang, Q., and Shao, J. (2019). Semantic trajectory compression via multi-resolution synchronization-based clustering. *Knowledge-Based Systems*, 174:177–193.

Hesabi, Z. R., Tari, Z., Goscinski, A., Fahad, A., Khalil, I., and Queiroz, C. (2015). Data summarization techniques for big data—a survey. In Khan, S. U. and Zomaya, A. Y., editors, *Handbook on Data Centers*, pages 1109–1152. Springer, New York, United States.

Lee, J.-G., Han, J., and Whang, K.-Y. (2007). Trajectory clustering: A partition-and-group framework. In *SIGMOD*, page 593–604, New York, NY, USA. ACM.

Machado, V. L., Mello, R. d. S., and Bogorny, V. (2022). A method for summarizing trajectories with multiple aspects. In *International Conference on Database and Expert Systems Applications, DEXA*, pages 433–446. Springer.

Machado, V. L., Mello, R. d. S., Bogorny, V., and Schreiner, G. A. (2024). A survey on the computation of representative trajectories. *GeoInformatica*, pages 1–26.

Machado, V. L. *et al.* (2023a). A method for computing representative data for multiple aspect trajectories based on data summarization. In *XXIV Brazilian Symposium on Geoinformatics*, GEOINFO.

Machado, V. L. *et al.* (2023b). Towards a representativeness measure for summarized trajectories with multiple aspects. In *XXIV Brazilian Symposium on Geoinformatics*, GEOINFO.

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

Panagiotakis, C., Pelekis, N., and Kopanakis, I. (2009). Trajectory voting and classification based on spatiotemporal similarity in moving object databases. In *International Symposium on Intelligent Data Analysis*, pages 131–142. Springer.

Petry, L. M. *et al.* (2019). Towards semantic-aware multiple-aspect trajectory similarity measuring. *Transactions in GIS*, 23(5):960–975.

Pugliese, C., Lettich, F., Pinelli, F., and Renso, C. (2023). Summarizing trajectories using semantically enriched geographical context. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–10.

Seep, J. and Vahrenhold, J. (2019). Inferring semantically enriched representative trajectories. In *1st ACM SIGSPATIAL Int. Workshop on Computing with Multifaceted Movement Data*, MOVE'19, pages 1–4, New York, NY, USA. ACM.

Tortelli Portela, T., Tyska Carvalho, J., and Bogorny, V. (2022). Hipermovelets: high-performance movelet extraction for trajectory classification. *International Journal of Geographical Information Science*, 36(5):1012–1036.

Varlamis, I., Kontopoulos, I., Tserpes, K., Etemad, M., Soares, A., and Matwin, S. (2021). Building navigation networks from multi-vessel trajectory data. *GeoInformatica*, 25:69–97.

Varlamis, I., Tserpes, K., Etemad, M., Júnior, A. S., and Matwin, S. (2019). A network abstraction of multi-vessel trajectory data for detecting anomalies. In *EDBT/ICDT Workshops*, volume 2019.

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