Journal of Information and Data Management, 2025, 16:1, doi: 10.5753/jidm.2025.4082

A Robust Measure for Evaluating Representativeness of Summarized Trajectories with Multiple Aspects

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***]
Lucas Vanini [**Instituto Federal Sul-Rio-Grandense (IFSUL)** | ***lucasvanini@ifsul.edu.br***]
Chiara Renso [**Consiglio Nazionale delle Ricerche (CNR) -ISTI** | ***chiara.renso@isti.cnr.it***]
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: 31 January 2024 • **Published:** 14 January 2025

Abstract As trajectory datasets grow larger, summarization techniques become increasingly important. However, current methods often lack a suitable measure of representativeness, making evaluation a complex task. This is especially true in the context of multi-aspect trajectories, where evaluating summarization techniques is particularly challenging. To address this, we have developed a novel representativeness measure called *RMMAT*. This innovative method combines similarity metrics and covered information, offering adaptability to diverse data and analysis needs. With *RMMAT*, evaluating summarization techniques is simplified, and deeper insights can be gained from extensive trajectory data. Our evaluation of real-world trajectory datasets demonstrates that *RMMAT* is a robust Representativeness Measure for Summarized Trajectories with Multiple Aspects. This measure could help researchers and analysts to evaluate and empower them to make informed decisions about the quality and relevance of representative data for their analytical goals.

Keywords: Multiple aspect trajectory, representative trajectory, trajectory similarity, representativeness measure

1 Introduction

The need to distill valuable insights is paramount in an era of vast trajectory data generated by individuals, vehicles, and objects. The proliferation of the Internet of Things (IoT) further enriches trajectories with multiple aspects, such as weather conditions during travel, the individual's mood, and social media posts. Extracting representative information from trajectories is crucial for effective analysis.

Trajectory summarization methods provide essential tools for creating concise representations, allowing analysts to comprehend and leverage the underlying movement patterns efficiently. Nevertheless, evaluating the effectiveness of these summarization techniques is a complex task, often hampered by the lack of a robust and comprehensive measure of representativeness [Seep and Vahrenhold, 2019; Machado *et al.*, 2022].

This article introduces the *Representativeness Measure for Multiple-Aspect Trajectories (RMMAT)*, addressing the challenge of assessing how well a representative trajectory reflects the original data. By applying the power of similarity metrics and covered information, *RMMAT* provides a multifaceted measure that quantifies the quality of representative trajectories in terms of their representativeness to the complete input dataset. This score, adaptable within a customizable configuration, empowers analysts to tailor the evaluation process to align the unique demands of their analytical scenarios.

By filling the void left by the lack of a comprehensive representativeness measure, *RMMAT* equips researchers with a potent tool for extracting insights from summarized *multiple-aspect trajectory (MAT)* data in the burgeoning trajectory data landscape. To help understand the paper better, the main acronyms used in the paper are summarized in Table 1.

In subsequent sections, we delve into *RMMAT*'s formulation, rigorous experimental evaluations, and facets related to similarity and covered information. We evaluate *RMMAT* using the *Foursquare* dataset (193 users), with promising results.

This paper is an extended version of Machado *et al.* [2023b], 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. Moreover, we have extended Section 5 by including a comparative experimental evaluation between two state-of-the-art MAT summarization methods, *MAT-SG* and *MAT-SGT*. Both methods establish a mapping between the input data and the resultant representative trajectory. With this comparative analysis, we aim to analyze the use of the covered information component in *RMMAT* measure and compare summarization methods. This analysis will give us insights into the computation of representative data and help us understand the differences between summarization methods.

The rest of this paper is organized as follows. Section 2 introduces foundational concepts. Section 3 is dedicated to problem and scope definition. Section 4 describes the proposed measure. Section 5 presents evaluations, and Section 6 concludes the paper.

Table 1. Acronyms table

Acronym	**Explanation**
AR	Average Recall
FSM	Finite State Machine
MAT	Multiple-Aspect Trajectory
MAT-SG	MAT Summarization based on a spatial Grid
MAT-SGT	MAT Summarization based on a spatial Grid and Temporal sequence
POI	Point of Interest
RMMAT	Representativeness Measure for MAT
RT	Representative Trajectory

2 Fundamentals

Trajectory data has become increasingly important in various fields due to the overall adoption of geolocation technologies. One of the foundational pillars of this work is the comprehensive exploration of moving objects. In data analytics, trajectory data is essential for mining, analysis, and decision-making, as it is being collected more frequently [Renso *et al.*, 2003; Oladimeji *et al.*, 2023].

The concept of a trajectory has evolved over time. Initially, a *raw trajectory* referred to the sequential movements of an object through geographical space over time, as defined by Erwig *et al.* [1999]. This raw trajectory comprised two dimensions: spatial and temporal. Around 2007, the notion of a *semantic trajectory* emerged. A third dimension was added, enriching the raw spatiotemporal trajectory *(x, y, t)* with semantic data. One example could be a *point of interest (POI)*, like a restaurant, that the object had visited [Alvares *et al.*, 2007; Parent *et al.*, 2013].

With the proliferation of the Internet of Things (IoT) and social media, trajectories have been further enriched with diverse semantic information. When trajectories, or their specific points, are associated with multiple semantic contexts, they are referred to as *multiple aspect trajectories (MAT)* [Mello *et al.*, 2019]. This trajectory also encompasses three dimensions (spatial, temporal, and semantic), but the semantic dimension can represent multiple and heterogeneous aspects.

As depicted in Figure 1, an individual's trajectory throughout a day serves as an example. The raw trajectory retains spatiotemporal data about the individual (Figure 1(a)). Conversely, Figure 1(b) illustrates a semantic trajectory, where contextual information is associated with the raw data, like PoIs (home, work, and restaurant).

Figure 1(c), in turn, showcases a raw trajectory enriched with multiple information, like the mean of transportation used by the individual, postings on social networks, weather conditions, health information, and so on. It emphasizes the complexity of MATs since the three dimensions can hold simple or complex attributes depending on the domain context. Moreover, MATs can generate vast amounts of data at high frequency, making it challenging to extract meaningful insights. In order to address this issue, a promising strategy is to compute summarized data from a set of MATs, as proposed in some works [Seep and Vahrenhold, 2019; Machado *et al.*, 2022, 2023a].

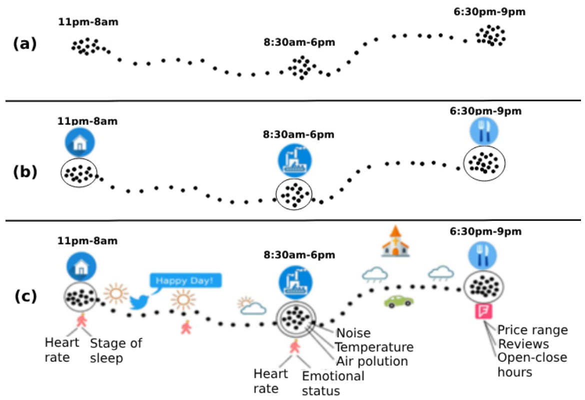

Figure 1. An example of a raw trajectory (a), semantic trajectory (b), and multiple aspect trajectory(c). Adapted from Mello *et al.* [2019].

2.1 Trajectory data summarization

Managing trajectory data is a big challenge due to the vast volume and variety of data continuously generated by different devices, resulting in an overwhelming volume and diversity of information [Martinez *et al.*, 2018; Gao *et al.*, 2019]. To address this issue, *Trajectory data summarization* is a vital process that condenses extensive and complex trajectories into more manageable and informative summaries [Etienne *et al.*, 2016].

Trajectory summarization aims at reducing the volume of trajectory data while preserving its essential characteristics and patterns in a more compact representation [Gao *et al.*, 2015]. In short, *MAT summarization* encompasses a process of abstraction from a set of MATs, culminating in a *representative MAT*.

The concept of representative trajectory is essential in trajectory summarization. According to Lee *et al.* [2007]; Ayhan and Samet [2015], a representative trajectory can be described as an imaginary trajectory that denotes the main behavior of a cluster of trajectories. Alternatively, Panagiotakis *et al.* [2012] suggests that a representative trajectory can vary according to the considered focus, like interest, density, frequency, and pairwise distance. It is worth noting that the representative MAT does not need to be congruent with every individual MAT, but it captures the overarching essence of the dataset [Machado *et al.*, 2022].

Representative trajectories are a useful way to analyze and visualize a dataset of trajectories. They help data analysts understand patterns in the data, which can be used to make better decisions. These patterns can serve as invaluable tools for diverse applications, such as analyzing traffic patterns within a city or identifying regions with elevated crime rates. As depicted in Figure 2 (left), the MATs across distinct days can give insights into an individual's movements. Meanwhile, the right side illustrates the culmination of these MATs into a representative MAT. This summarized representation effectively encapsulates the individual's frequent activities.

As shown in Figure 2 (right), the system can recognize a pattern in an individual's preference for lunch in a vegetarian restaurant. Therefore, if the user is in a different location than their usual lunch place around lunchtime, the system

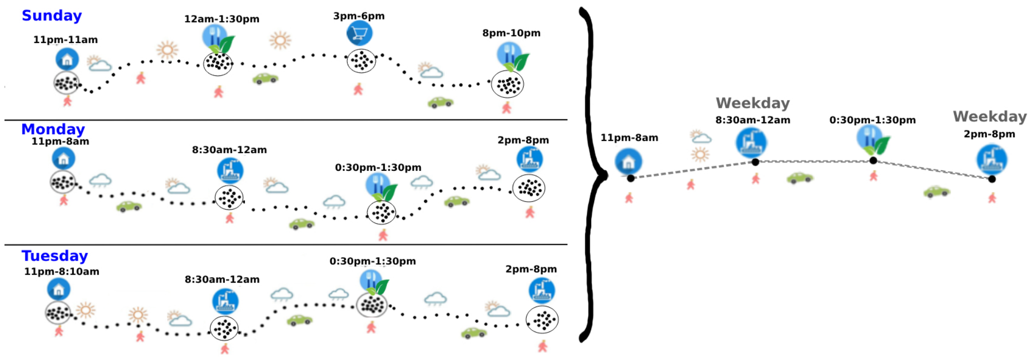

Figure 2. Examples of MATs (left) and a representative MAT for them (right) [Machado *et al.*, 2022].

can identify vegetarian restaurants nearby and recommend them to the user.

Trajectory summarization is pivotal in handling and extracting insights from trajectory data. It reduces data complexity while preserving essential information for various applications. The summarized representation of a moving object can help to understand its behavior easily.

Some methods have been proposed for computing a *Representative Trajectory (RT)* from a set of MAT, focusing on movement patterns. In 2019, Seep and Vahrenhold [2019] proposed a Finite State Machine (FSM) to identify common transitions among movements yielding the *RT* by representing each state as a common point. The sequence of states yields the *RT*. However, this method did not consider the aspect-specific types within MATs since all attributes of the MAT points are spatial or non-spatial. In 2022, a method designed explicitly for MATs was proposed (*MAT-SG*) [Machado *et al.*, 2022]. *MAT-SG* segments the input MATs into a spatial grid and performs summarization within each relevant cell, treating all aspects of MATs individually. Recently, in 2023, *MAT-SGT* was proposed as a data summarization method specifically designed to compute representative MATs identifying the temporal sequence associated with the movement pattern. It segments the input MAT points in two steps: a spatial grid and significant temporal intervals, summarizing all information. Both *MAT-SG* and *MAT-SGT* establish a mapping between the input MATs and the computed representative MAT, preserving the relationship between the original data and its summarized representation.

3 Problem Definition

In Figure 2, an example of trajectory summarization applied to input dataset **D** ($\boldsymbol{D} = \{p, q, r\}$) generates the *RT*). However, an issue with existing literature is the lack of a well-defined measure for evaluating how well the representative data accurately represents the entire dataset **D**. Studies highlight this common challenge when computing representative trajectories from MATs [Seep and Vahrenhold, 2019; Machado *et al.*, 2022, 2023a].

This paper intends to answer this fundamental question: 'How much of the *RT* captures and reflects the original MATs' essence within an input dataset **D**?'. The computation of *RT*'s should align with specific use case objectives and requirements, as different applications may necessitate varying levels of granularity and information preservation [Machado *et al.*, 2022].

The scope of this work is to propose a novel representativeness measure tailored for big trajectory data with multiple aspects, aiming to quantify how much information the *RT* covers from the input dataset **D** and how similar this *RT* is to the entire dataset, i.e., it aims to measure how well a representative trajectory captures the essence of the original dataset, which is particularly useful given the increasing complexity and growth of trajectory data. The objective is to simplify the evaluation of summarization methods and extract valuable insights from extensive MAT datasets.

4 RMMAT: Representativeness Measure for Multiple-Aspect Trajectory

In this section, we present the core concepts of our work, known as the *Representativeness Measure for Multiple-Aspect Trajectories* (*RMMAT*), which serves as a standardized metric for assessing the effectiveness of representative data produced by summarization methods. *RMMAT* considers similarity metrics and covered information to measure the quality of representative data. It is specifically designed to fit different analytical scenarios, enabling analysts to tailor the evaluation process to their specific needs.

RMMAT provides a concrete and measurable way to assess the quality of representative data by considering similarity and covered information. It offers valuable insights into the summarization quality, allowing for a rigorous and objective evaluation of how well the representative data captures the intricacies of the data. By introducing *RMMAT*, we aim to simplify the evaluation of summarization methods and extract valuable insights from extensive MAT datasets.

RMMAT consists of two integral components: (i) similarity metric and (ii) covered information, aiming to quantify the information coverage of *RT* from the input dataset **D** and estimate its similarity to the entire dataset. By combining these two components, *RMMAT* aims to overcome the limitations of evaluating representativeness in summarized MAT. In the following, we delve into the details of each component.

4.1 Similarity Metric Component

The trajectory similarity metric measures how similar two trajectories are based on attributes such as spatial positions, temporal sequences, and potentially additional semantic aspects. It quantifies how much they share common patterns in terms of movement through space, time, and semantics. While traditional measures compare trajectories pairwise, the challenge is to measure the similarity of an *RT* against the entire dataset of trajectories.

We calculate the similarity between *RT* and each $\{T_1, T_2, ..., T_n\} \in D$, considering that D and *RT* are non-empty. We use the median value of the similarity measure to account for skewed distributions or outliers in the dataset. To address this concern, we opt to use the median value of the similarity measure across all pairs of MATs (RT and each $T \in D$), given that $0 \leq Similarity \leq 1$. The median is less affected by extreme values or anomalies in similarity scores, resulting in a more balanced representation of central tendency. The equation is given by:

$$\begin{aligned} &|\text{Similarity}\,(RT, D)| = \\ &Me(\{\text{Similarity}(RT, T_1), \text{Similarity}(RT, T_2), \ldots, \text{Similarity}(RT, T_n)\}) \end{aligned} \tag{1}$$

Find the median similarity value between RT and all $T \in D$ by using the function Me that calculates the median of similarity scores.

4.2 Covered Information Component

In order to compute the covered information within D by RT, we evaluate the MAT points of each $T_i \in D$ that RT covers and aim to derive the proportion of covered information in a non-negative value. This computation is defined as:

$$\left(\frac{\sum_{p \in T}^{D} p \subseteq RT}{|D.points|}\right) \tag{2}$$

The objective of *RMMAT* is to harmonize both components: (i) the similarity between RT and all MATs and (ii) the measure of the coverage input MAT points by RT, when available. So, the representativeness measure score between the RT and the input dataset is calculated by the final function *RMMAT*, with *RMMAT* ∈ [0,1]:

$$\begin{aligned} \mathit{RMMAT} = &\omega_{sim} \times |\text{Similarity}\,(RT, D)| \\ &+ \omega_{cover} \times \left(\frac{\sum_{p \in T}^{D} p \subseteq RT}{|D.\text{points}|}\right) \end{aligned} \tag{3}$$

Let $W = \{\omega_{sim}, \omega_{cover}\}$ a non-empty set of the weights. The weights ω_{sim} and ω_{cover} represent the importance of each component for computing the representativeness between trajectories for a specific scenario. We assume that $\omega_{sim} + \omega_{cover} = 1.0$. Components with higher weights have a more pronounced impact on the final representativeness scores.

5 Experimental Evaluation

In this section, we illustrate the functioning of *RMMAT* through a practical example and evaluate its performance using experimentation on a real dataset. This assessment gauges its accuracy and practical utility in effectively capturing trajectory data. The experiments were conducted on a Dell Inspiron laptop with an Intel Core i5 processor and 16 GB memory implemented with Java. We describe the datasets (Section 5.1), the general experimental setup (Section 5.2), and two evaluations analyzing the relevance of RT concerning similarity information and covered information (Sections 5.3 and 5.4) in the following sections.

5.1 Dataset

We used the Foursquare NYC dataset, which includes check-in records from April 2012 to February 2013 in New York City. The dataset is enriched with contextual information such as *weekday*, *category*, *price*, *rating* of the POIs, and *weather conditions*. The dataset includes 3079 trajectories from 193 users, with each trajectory containing around 22 data points, and each user is associated with an average of about 16 trajectories.

5.2 General Experimental Setup

To compute *RMMAT*, several crucial elements need to be defined. First, we need to select a summarization method to derive representative data. In this work, we use the state-of-the-art MAT summarization methods *MAT-SG* and *MAT-SGT*. These methods establish a mapping between the input data and the resultant representative trajectory, as exemplified in Section 2, which allows us to include covered information in the computation of representativeness. Second, an appropriate similarity measure is needed, and we use the widely recognized MAT similarity measure called *MUITAS* [Petry *et al.*, 2019] to establish trajectory similarity. Finally, we define weights (W) to individual components using a balanced weights strategy, setting $\omega_{sim} = \omega_{cover} = \frac{1}{2}$.

5.2.1 Summarization method setup

Both *MAT-SG* and *MAT-SGT* summarize data on a grid of cells. Two parameters are required for its setup: (i) τ_{rv} (threshold RV), which determines representative values, and (ii) τ_{rc} (threshold RC), which sets the minimum number of MAT points for a cell to qualify for summarization. For evaluation purposes, we dynamically define the cell size of the spatial grid for *MAT-SG* using the same strategy of *MAT-SGT*. This was achieved by iteratively analyzing the representative trajectory for different values of z and selecting the optimal value that yields the best *RT*.

We performed experiments by executing *MAT-SG* and *MAT-SGT* in each ground truth, i.e., we consider each user as the criterion to cluster MATs into groups. The method was repeated for each user with different parameter settings (τ_{rv} and τ_{rc}), varying from 0% to 25% (0, 1, 5, 10, 15, 20, 25), resulting in 49 runs for each user. This parameter variation allows for evaluating the sensitivity and robustness of the *RMMAT* measure.

We established our criteria since we did not identify a common strategy to evaluate a representative MAT to be used as a benchmark in the existing literature. For each group, we select the MAT T_i with the median similarity score across all trajectories in the group, i.e., the i^{th} MAT $T \in D$ that satisfies the criterion of the median similarity across all the others in D. This ensures that the baseline acts as a reference point for comparison purposes.

5.2.2 Similarity Measure setup

We performed the similarity measure with MUITAS, where settings must be defined, including features, weight, and proximity functions. Each attribute in the input dataset is defined as a single feature. Proximity functions consider spatial, temporal, and semantic aspects with weight-balanced dimensions. The formats of a simple MAT and a *RT* created by *MAT-SG* and *MAT-SGT* are different, requiring special analysis and settings. In a simple MAT, each attribute has only one value. In contrast, in *RT*, rank values may be present, particularly when handling categorical values in the semantic dimension and the temporal dimension. Consequently, particular analyses and configurations become essential.

Adopted functions are:

(i) for the spatial dimension, we use the *Euclidean distance*. A spatial match occurs if the distance between a trajectory T_j in the group and *RT* coordinates is within four times the *RT* computed threshold, which is determined by the spatial dispersion of the MAT points in both *MAT-SG* and *MAT-SGT*;

(ii) for the temporal dimension, we assess the match between RT and other trajectories T_j in the group by evaluating the *temporal interval* of *RT*. A match occurs if the timestamp of T_j lies within the interval. The baseline, which follows the same format as input trajectories, uses a 30, 45, or 60-minute threshold for analysis;

(iii) for semantic dimension, we evaluate attribute matching for *numeric* and *categorical* data types. For numeric data types, a match occurs if the difference in attribute values is $<= 10\%$ of the *RT* value. For categorical data types, a match occurs if the attribute value falls within the range of *RT* values.

For the sake of understanding, this section introduces a Running Example to illustrate the functionality of *RMMAT*. It consists of a set of input MATs **D**, each one representing a trajectory attributed to a different individual.

The input MATs and their corresponding *RT* are shown in Figure 3. The trajectories are depicted on the left side, and their corresponding *RT* calculated is shown on the right side. The spatial and temporal information, along with the price and category of the PoIs, weather conditions, and precipitation, represent the input trajectories and the *RT*.

For computing *RMMAT*, we first compute the similarity between each trajectory in **D** and *RT*, where $MUITAS(q, RT) = 0.686$, $MUITAS(r, RT) = 0.835$, and $MUITAS(s, RT) = 0.871$. Then, according to Equation 1, the $|Similarity\,(RT, D)| = 0.835$. Regarding the covered information, Equation 2, $T^c(RT) = \frac{10}{17} = 0.5882$.

Finally, considering the computation of *RMMAT* with balanced weights strategy by setting $\omega_{sim} = \omega_{cover} = \frac{1}{2}$, and according to Equation 3, we have $RMMAT = (0.5 \times 0.835) + (0.5 \times 0.5882) = 0.7116$. It means that *RT* has a representativeness of 0.7116 of **D** considering both similarity and covered information. While the similarity score with input MATs is 0.835, the impact of its coverage of data points is 0.5882. Therefore, in the context of both similarity and coverage, it can be said that the representativeness of *RT* has an impact of approximately 71% of the input MATs.

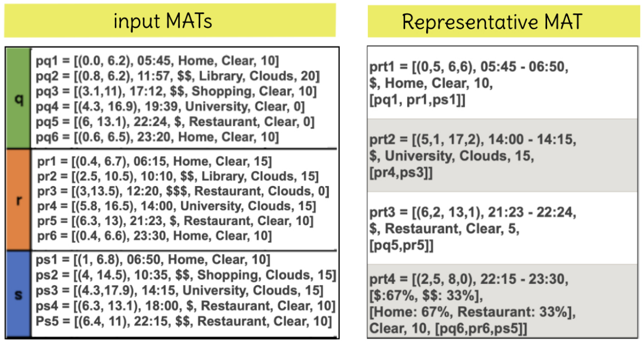

Figure 3. Set of input MATs $\mathbf{D} = \langle q, r, s \rangle$, where $q = \langle p_{q_1}, p_{q_2}, ..., p_{q_n} \rangle$, $r = \langle p_{r_1}, p_{r_2}, ..., p_{r_m} \rangle$, and $s = \langle p_{s_1}, p_{s_2}, ..., p_{s_t} \rangle$ (left), and their correspondent *RT* (right).

5.3 Analyzing *RMMAT* Regarding Similarity Information

To gain insights into *RMMAT* behavior, we conducted an experiment using a sample of user trajectories of the Foursquare dataset. We presented illustrative examples of evaluations based on the standard deviation (SD) of average and median similarity scores of each user's baseline. We selected three users for analysis: (i) user 185 ($\mathbf{D}_{u185}$), which has a lower SD for average similarity scores; (ii) user 730 ($\mathbf{D}_{u730}$), which has a lower SD for median similarity scores; and (iii) user 708 ($\mathbf{D}_{u708}$), showcasing the highest SD for both average and median similarity scores.

This experiment analyzes the representativeness of RTs in similarity information with different threshold values for relevant cell (RC) and representativeness value (RV), namely τ_{rc} and τ_{rv}, using $\omega_{sim} = 1$ and $\omega_{cover} = 0$ based on MUITAS. The investigation explores the impact of varying combinations of these thresholds on the computation of *RT* in both *MAT-SG* and *MAT-SGT*.

Figures 4 and 5 visually depict the results of the similarity evaluation for each user under different input parameter configurations, compared to the baseline. These figures highlight the variations in similarity scores while varying the temporal threshold.

Our representativeness measure consistently outperformed the baseline for low parameter configurations, shedding light on the intricate interplay between different threshold parameters and their impact on *RT* computed from MUITAS.

For *MAT-SG*, users 185 and 708 exhibit a specific *RT* behavior pattern across different τ_{rv} values. Regarding the τ_{rc}, determining relevant cells for *RT* computation seems to influence *RT* changes significantly since, for these users, an increase in the value of this parameter configuration results in a decrease in *RMMAT*. This underscores the sensitivity of *RMMAT* to parameter choices and their implications for the representativeness of *RT*. The behavior of user 730 highlights the importance of parameter configurations in *RT* com-

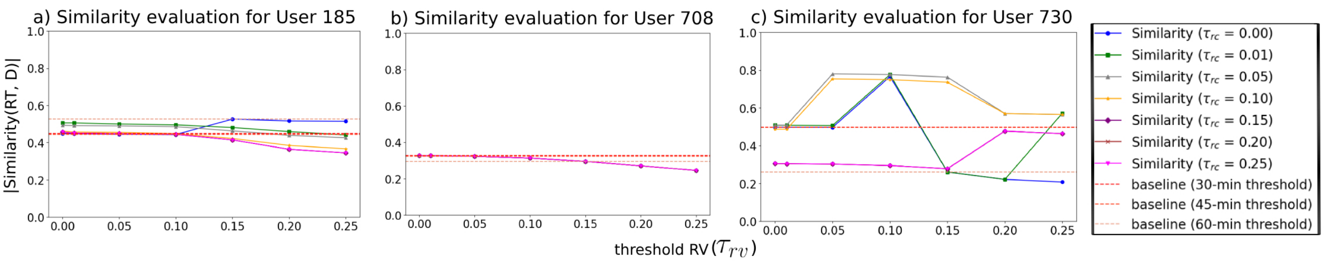

Figure 4. This graph analyzes the similarity evaluation (Y-axis) by comparing varying threshold RC, the τ_{rc}, shown as distinct lines, and the threshold RV, the τ_{rv}, concerning baseline for users 185, 708, and 730. It explores different parameter configurations of the τ_{rv} (X-axis) to evaluate similarity. This analysis refers to the *MAT-SG* method.

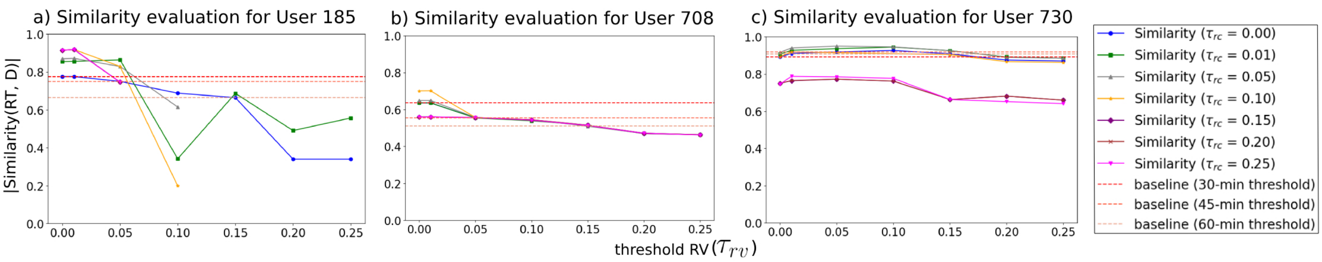

Figure 5. This graph analyzes the similarity evaluation (Y-axis) by comparing varying threshold RC, the τ_{rc}, shown as distinct lines, and the threshold RV, the τ_{rv}, concerning baseline for users 185, 708, and 730. It explores different parameter configurations of the τ_{rv} (X-axis) to evaluate similarity. This analysis refers to the *MAT-SGT* method.

putation.

For *MAT-SGT*, users 708 and 730 display specific *RT* behavior patterns across different τ_{rv} values. As the value of this parameter configuration increases, *RMMAT* decreases, emphasizing the influence of parameter configurations on *RT* computation and its subsequent impact on representativeness.

We employed correlation coefficients to quantify the impact of values for τ_{rc} and τ_{rv} in both methods on the *RMMAT* measure. The coefficients reveal relationships between input parameters and *RMMAT* scores for *RT* computed for both methods (*MAT-SG* and *MAT-SGT*) and input trajectories. The results in Table 2 offer valuable insights into how threshold parameters influence the accuracy of computed representative trajectories. Positive coefficients indicate that higher threshold values correspond to higher *RMMAT* scores, while negative coefficients suggest the opposite.

Table 2. Impact of Input Parameters on the Representativeness Measure of *RT*

correlation coefficient	***MAT-SG***		***MAT-SGT***	
	τ_{rc}	τ_{rv}	τ_{rc}	τ_{rv}
User 185	-0.568	-0.526	0.408	-0.788
User 708	-8.770	-0.966	-0.154	-0.829
User 730	-0.378	0.027	-0.817	-0.243

For *MAT-SG*, user 185 exhibits a negative correlation (-0.568) between *RMMAT* scores and τ_{rc}, indicating that increasing τ_{rc} leads to a decrease in *RMMAT* scores. User 708, characterized by a greater SD in similarity scores and displayed the one with a more consistent pattern, shows a high negative correlation (-8.770), suggesting that higher τ_{rc} values consistently lead to lower *RMMAT* scores. For user 730, a negative correlation (-0.378) implies that higher τ_{rc} values result in lower *RMMAT* scores. Across all users in *MAT-SG*, the negative correlation pattern highlights that higher τ_{rc} values lead to less representative *RT*.

For *MAT-SGT*, user 185 exhibits a positive correlation (0.408) between *RMMAT* scores and τ_{rc}. The *RMMAT* scores increase as τ_{rc} values increase. User 708, characterized by greater SD in similarity scores, shows a slight negative correlation (-0.154), indicating that increasing τ_{rc} leads to a minor decrease in *RMMAT* scores. For user 730, who displays more consistent patterns, a negative correlation (-0.817) suggests that higher τ_{rc} values lead to lower *RMMAT* scores.

This analysis provides nuanced insights into the dynamics of *RMMAT* concerning similarity information. It comprehensively explains how different parameter configurations influence the computed *RT* and its representativeness. Notably, in *MAT-SG*, higher τ_{rv} values consistently lead to less representative *RT*. Meanwhile, in *MAT-SGT*, the correlation patterns reveal the nuanced impact of both τ_{rc} and τ_{rv} values on *RMMAT* scores. The parameter configuration significantly influences the behavior and accuracy of the computed representative trajectory, necessitating careful consideration of their selection to capture relevant input data patterns. This analysis underscores the improvements achieved through the *RMMAT* measure, highlighting its efficacy in enhancing data comprehension. Overall, the results emphasize the effectiveness of *RMMAT* as a valuable tool for better understanding complex trajectory data.

5.4 Analyzing *RMMAT* Regarding Covered Information

In the absence of a standardized strategy for evaluating the representativeness of a representative MAT in the existing literature, our analysis extends beyond similarity to encompass both similarity and cover components. To gauge the utility of *RT*, we employ the *Average Recall (AR)* metric, drawing inspiration from the experimental evaluation of the similarity measure proposed by Petry *et al*. [2019]. While aligning with their evaluation methodology and leveraging their dataset for ground truth segmentation, our focus diverges. In Petry *et al*. [2019], the primary objective was to validate their similarity measure, explicitly assessing the similarity between pairs of trajectories. While our foundation is rooted in their methodology, our focus remains to quantify the quality of the summarization methods and representativeness of data computation, evaluating the utility of *RT* within the context of the input dataset. We aim to evaluate the utility of *RT* within the context of the input dataset.

The AR metric becomes pivotal in this evaluation. This metric measures recall based on the similarity between the *RT* computed by *RMMAT* and other trajectories within the dataset. The recall is defined as the fraction of relevant trajectories that are successfully retrieved. In the context of ranking trajectories within the same ground truth group, the ideal outcome is that the top k most similar trajectories also belong to the same group, where $k = |\mathbf{D}|$. This provides a robust measure of how effectively *RT* can rank trajectories within the same group.

The evaluation process involves computing the *RT* for each user in our sample of users in our selected sample (users 185, 708, and 730), i.e., $\mathbb{T} = \{\mathbf{D}_{u185}, \mathbf{D}_{u708}, \mathbf{D}_{u730}, \cdots\}$. The idea is that the trajectories of the same user exhibit similarity. The goal is for each user of the *RT* to have high similarity values with the trajectories in that group.

To analyze the impact of covered information in *RMMAT*, we assess the utility of *RT* using the AR metric. The process begins by computing *RT* and calculating similarity over the entire dataset. Trajectories are then ordered based on similarity scores. Subsequently, trajectories are ranked according to these similarity scores, and the recall metric is computed. This metric quantifies how effectively *RT* can accurately rank trajectories within the same group.

Tables 3a and 3b display the AR values for user 185 by *MAT-SG* in scenarios without and with covered information, respectively. Additionally, Table 3c consolidates the outcomes of the AR analysis, indicating consistent results for both scenarios in this specific user context. Instances with missing values, indicated by ”-”, denote situations where *RT* computation with specific parameter configurations is not feasible due to the particular data patterns present in the input dataset. The variations between both methods are highlighted, with the higher value between with or without covered information being underlined.

Table 3. The AR Analysis of User 185 by *MAT-SG*

(a) AR without covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	0.9	1	1	1	1	1	1
0.01	0.9	1	1	1	1	1	1
0.05	0.9	1	1	0.95	0.95	0.95	0.95
0.10	0.9	1	1	1	1	1	1
0.15	0.9	0.98	1	1	1	1	1
0.20	0.9	1	1	1	1	1	1
0.25	0.9	0.98	1	1	1	1	1

(b) AR with covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	0.9	1	1	1	1	1	1
0.01	0.9	1	1	1	1	1	1
0.05	0.9	1	1	0.95	0.95	0.95	0.95
0.10	0.9	1	1	1	1	1	1
0.15	0.9	0.98	1	1	1	1	1
0.20	0.9	1	1	1	1	1	1
0.25	0.93	0.98	1	1	1	1	1

(c) AR Analysis regarding covered information

	With Cover	**Without Cover**
Missing values	0	0
Best Value	1	1
Worse Value	0.9	0.9
AVG AR	0.988	0.988
Median AR	1	1

Table 4. The AR Analysis of User 185 by *MAT-SGT*

(a) AR without covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	0.9	0.93	0.95	1	1	1	1
0.01	0.9	0.93	0.93	1	1	1	1
0.05	0.9	0.95	0.98	1	1	0.98	0.98
0.10	0	0	0.81	0	-	-	-
0.15	0	0.98	-	-	-	-	-
0.20	0.02	1	-	-	-	-	-
0.25	0.02	0.83	-	-	-	-	-

(b) AR with covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	0.9	0.93	0.95	1	1	1	1
0.01	0.9	0.93	0.93	1	1	1	1
0.05	0.9	0.95	0.98	1	0.98	0.98	0.98
0.10	0	0	0.81	0	-	-	-
0.15	0	0.98	-	-	-	-	-
0.20	0.02	1	-	-	-	-	-
0.25	0.02	0.83	-	-	-	-	-

(c) AR Analysis regarding covered information

	With Cover	**Without Cover**
Missing values	18	18
Best Value	1	1
Worse Value	0	0
AVG AR	0.771	0.707
Median AR	0.93	0.93

For the same user 185, the scenarios for *MAT-SGT* are respectively presented in Tables 4a and 4b, and the compiled results of the AR analysis are presented in Table 4c.

Upon examining the summarized outcomes of the AR analysis in Table 4c, notable variations between scenarios that include and exclude covered information for User 185 by *MAT-SGT* become evident. Specifically, there is an average AR growth of 0.707 when analyzing the scenario without covered information, compared to 0.771 when including covered information.

In the case of User 708, computed by *MAT-SG*, Tables 5a and 5a show the AR values, and Table 5c compiles the results of the AR analysis, where for this situation, both scenarios present the same results.

By *MAT-SGT*, both scenarios for user 708 are respectively presented in Tables 6a and 6b, and the compiled results of the AR analysis are presented in Table 6c. While there were some minor variations in the specific values, the overall assessment presented in Table 6c does not indicate a substantial difference. The AR values for this user are relatively stable, regardless of whether the covered information was included or excluded during the analysis.

For the user 730, computed by *MAT-SG*, both scenarios are respectively presented in Tables 7a and 7b, and the compiled results of the AR analysis are presented in Table 7c. A slight

Table 5. The AR Analysis of User 708 by *MAT-SG*

(a) AR without covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	1	1	0.9	0.8	0.5	0.5	0.5
0.01	1	1	0.9	0.8	0.5	0.5	0.5
0.05	1	1	0.9	0.8	0.6	0.6	0.6
0.10	1	1	0.9	0.8	0.6	0.6	0.6
0.15	1	1	0.9	0.7	0.7	0.7	0.5
0.20	1	1	0.9	0.7	0.7	0.7	0.6
0.25	1	1	0.9	0.8	0.6	0.6	0.6

(b) AR with covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	1	1	0.9	0.8	0.5	0.5	0.5
0.01	1	1	0.9	0.8	0.5	0.5	0.5
0.05	1	1	0.9	0.8	0.6	0.6	0.6
0.10	1	1	0.9	0.8	0.6	0.6	0.6
0.15	1	1	0.9	0.7	0.7	0.7	0.5
0.20	1	1	0.9	0.7	0.7	0.7	0.6
0.25	1	1	0.9	0.8	0.6	0.6	0.6

(c) AR Analysis regarding covered information

	With Cover	**Without Cover**
Missing values	0	0
Best Value	1	1
Worse Value	0.5	0.5
AVG AR	0.81	0.81
Median AR	0.7	0.7

Table 6. The AR Analysis of User 708 by *MAT-SGT*

(a) AR without covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.01	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.05	0.8	0.8	0.8	0.8	0.8	0.8	0.8
0.10	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.15	0.8	0.8	0.8	0.8	0.8	0.8	0.9
0.20	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.25	0.9	0.9	0.9	0.9	0.8	0.8	0.8

(b) AR with covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	0.8	0.8	0.9	0.8	0.9	0.9	0.9
0.01	0.8	0.8	0.9	0.8	0.9	0.9	0.9
0.05	0.8	0.8	0.8	0.8	0.8	0.8	0.8
0.10	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.15	0.8	0.8	0.8	0.8	0.8	0.8	0.8
0.20	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.25	0.9	0.9	0.9	0.9	0.8	0.8	0.8

(c) AR Analysis regarding covered information

	With Cover	**Without Cover**
Missing values	0	0
Best Value	0.9	0.9
Worse Value	0.8	0.8
AVG AR	0.862	0.87
Median AR	0.9	0.9

Table 7. The AR Analysis of User 730 by *MAT-SG*

(a) AR without covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	1	1	1	1	0.9	0.83	0.83
0.01	1	1	1	1	0.93	0.87	0.87
0.05	1	1	1	1	0.93	0.87	0.87
0.10	1	1	1	1	0.93	0.87	0.87
0.15	1	1	1	1	0.9	0.83	0.83
0.20	1	1	1	1	0.93	0.87	0.87
0.25	1	1	1	1	0.93	0.87	0.87

(b) AR with covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	1	1	1	1	0.93	0.83	0.83
0.01	1	1	1	1	0.93	0.87	0.87
0.05	1	1	1	1	0.93	0.87	0.87
0.10	1	1	1	1	0.93	0.87	0.87
0.15	1	1	1	1	0.9	0.83	0.83
0.20	1	1	1	1	0.93	0.87	0.87
0.25	1	1	1	1	0.93	0.87	0.87

(c) AR Analysis regarding covered information

	With Cover	**Without Cover**
Missing values	0	0
Best Value	1	1
Worse Value	0.83	0.83
AVG AR	0.940	0.927
Median AR	1	1

variation can be observed in this situation when including or excluding covered information, showing in underlying value. Additionally, the average AR growth of 0.927 when analyzing the scenario without covered information, compared to 0.940 when including covered information.

The AR values for user 730 computed by *MAT-SGT* in both scenarios are presented in Tables 8a and 8b. Additionally, Table 8c compiles the AR analysis outcomes for this user. It is evident that there is a substantial variation in AR values across different scenarios, which highlights the significant impact of covered point data on the AR measure. This disparity emphasizes how the inclusion of covered information can significantly influence the outcomes of a representativeness measure.

Analyzing *RMMAT* regarding covered information and observing the variation in AR values between the inclusion and exclusion of covered point data reveals consistent trends in both *MAT-SG* and *MAT-SGT* scenarios. Overall, minimal differences are observed, suggesting a stable pattern of minimal variation. In the case of *MAT-SG*, there is a slight growth when covered information is included. Notably, User 730 in the *MAT-SGT* scenario exhibits the most significant distinctions between scenarios, emphasizing the influence of covered data points. However, it is intriguing to observe that, for

Table 8. The AR Analysis of User 730 by *MAT-SGT*

(a) AR without covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	0.97	0.97	0.9	0.9	0.9	0.9	0.9
0.01	0.93	0.93	0.87	0.87	0.87	0.87	0.87
0.05	0.93	0.93	0.87	0.87	0.87	0.87	0.87
0.10	0.97	0.97	0.83	0.83	0.83	0.83	0.83
0.15	0.9	0.9	0.77	0.77	0.77	0.77	0.77
0.20	0.9	0.9	0.83	0.83	0.83	0.83	0.83
0.25	0.87	0.87	0.83	0.83	0.83	0.83	0.83

(b) AR with covered information

τ_{rv} \ τ_{rc}	**0.00**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
0.00	1	1	1	1	0.9	0.9	0.87
0.01	1	1	1	1	0.93	0.93	0.87
0.05	1	1	1	1	0.9	0.9	0.87
0.10	1	1	1	1	0.87	0.87	0.83
0.15	1	1	1	1	0.9	0.9	0.73
0.20	1	1	1	1	0.87	0.87	0.9
0.25	1	1	1	1	0.93	0.93	0.87

(c) AR Analysis regarding covered information

	With Cover	**Without Cover**
Missing values	0	0
Best Value	1	0.97
Worse Value	0.73	0.77
AVG AR	0.94	0.878
Median AR	1	0.87

the same user, trajectories retrieved with covered data points fare better than computed RT trajectories, indicating a potential impact on *RMMAT* scores and implying differences in underlying data patterns.

In short, the AR analysis of User 708 by *MAT-SG* appears relatively unaffected by the presence of covered point data, indicating limited influence on the outcomes. In contrast, the analysis of User 730 by *MAT-SGT* underscores the substantial impact of aggregating covered information. This disparity underscores the importance of a nuanced consideration of each component in *RMMAT* computation, enabling a tailored parameter configuration to the specific dataset and analysis objectives.

6 Conclusion

In conclusion, this paper introduces the *Representativeness Measure for Multiple Aspect Trajectories (RMMAT)*, offering a standardized and comprehensive metric to assess the efficacy of representative data derived from trajectory summarization methods. As trajectory data experiences exponential growth and heightened complexity, *RMMAT* emerges as a pivotal tool for quantifying how well a representative trajectory encapsulates the essence of the original dataset.

Leveraging similarity metrics and covered information, *RMMAT* provides a holistic evaluation approach, enabling analysts to estimate both the similarity between representative and input trajectories and the information coverage within the dataset. This measure helps researchers and analysts evaluate and empowers them to make informed decisions about the quality and relevance of representative data for their analytical goals.

RMMAT effectively quantifies the representativeness of computed representative data in comparison to the original MATs, yielding valuable insights. For example, in evaluating *MAT-SG* and *MAT-SGT* methods, our findings underscore the pivotal role of parameter selection in achieving optimal results. This observation emphasizes how *RMMAT* offers insights that guide researchers in refining trajectory summarization methods for improved outcomes.

A notable strength of *RMMAT* lies in its adaptability, with configurable components that permit analysts to tailor the evaluation process to the unique demands of different analytical scenarios. This adaptability positions *RMMAT* as a versatile tool aligning with varying objectives and data characteristics.

Our work bridges a critical gap in the field of trajectory data summarization, allowing researchers and analysts to evaluate and measure trajectory summarization methods by a quantitative metric. By overcoming the limitations of previous subjective evaluation methods, *RMMAT* provides insights that guide researchers in refining trajectory summarization methods for improved outcomes.

As the effectiveness of computing an *RT* depends on the specific purpose and requirements of a use case, different applications may necessitate different levels of granularity and information preservation. The evaluation of this approach is inherently tied to the intended analytical objectives, focusing on views of similarity and covered information. Future work aims to explore additional perspectives regarding the representativeness of summarized MAT, such as reduced information.

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, Chiara, and Ronaldo contributed to defining the *RMMAT* measure, while Lucas 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 *RMMAT* is available at `https://github.com/RepresentantativeMAT/RMMAT.git`. The Foursquare NYC dataset used during the current experimental evaluation is available in `https://github.com/bigdata-ufsc/datasets`.

References

Alvares, L. O. *et al.* (2007). A model for enriching trajectories with semantic geographical information. In *15th ACM International Symposium on Geographic Information Systems, ACM-GIS*, page 22. ACM. DOI: 10.1145/1341012.1341041.

Ayhan, S. and Samet, H. (2015). Diclerge: Divide-cluster-merge framework for clustering aircraft trajectories. In *Proceedings of the 8th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, pages 7–14. DOI: 10.1145/2834882.2834887.

Erwig, M., Guting, R. H., Schneider, M., and Vazirgiannis, M. (1999). Spatio-temporal data types: An approach to modeling and querying moving objects in databases. *GeoInformatica*, 3(3):269–296. DOI: 10.1023/A:1009805532638.

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. DOI: 10.1080/13658816.2015.1081205.

Gao, C., Zhao, Y., Wu, R., Yang, Q., and Shao, J. (2015). Data summarization techniques for big data—a survey. In *Handbook on Data Centers*, pages 1109–1152. Springer. DOI: 10.1007/978-1-4939-2092-1_38.

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. DOI: 10.1016/j.knosys.2019.03.006.

Lee, J.-G., Han, J., and Whang, K.-Y. (2007). Trajectory clustering: A partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, page 593–604, New York, NY, USA. Association for Computing Machinery (ACM). DOI: 10.1145/1247480.1247546.

Machado, V., Portela, T., de Lara Machado, A., Schreiner, G., and dos Santos Mello, R. (2023a). A method for computing representative data for multiple aspect trajectories based on data summarization. *Proceedings of the Brazilian Symposium on GeoInformatics*, pages 25–36.

Machado, V., Portela, T., Renso, C., and dos Santos Mello, R. (2023b). Towards a representativeness measure for summarized trajectories with multiple aspects. *Proceedings of the Brazilian Symposium on GeoInformatics*, pages 37–48.

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. DOI: 10.1007/978-3-031-12423-5_33.

Martinez, D. *et al.* (2018). Smart data fusion: Probabilistic record linkage adapted to merge two trajectories from different sources. In *Eighth Sesar Innovation Days*.

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. DOI: 10.1111/tgis.12526.

Oladimeji, D., Gupta, K., Kose, N. A., Gundogan, K., Ge, L., and Liang, F. (2023). Smart transportation: an overview of technologies and applications. *Sensors*, 23(8):3880. DOI: 10.3390/s23083880.

Panagiotakis, C., Pelekis, N., Kopanakis, I., Ramasso, E., and Theodoridis, Y. (2012). Segmentation and sampling of moving object trajectories based on representativeness. *IEEE Transactions on Knowledge and Data Engineering*, 24(7):1328–1343. DOI: 10.1109/TKDE.2011.39.

Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., Macedo, J., Pelekis, N., *et al.* (2013). Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4):42:1–42:32. DOI: 10.1145/2501654.2501656.

Petry, L. M., Ferrero, C. A., Alvares, L. O., Renso, C., and Bogorny, V. (2019). Towards semantic-aware multiple-aspect trajectory similarity measuring. *Transactions in GIS*, 23(5):960–975. DOI: 10.1111/tgis.12542.

Renso, C., Spaccapietra, S., and Zimányi, E., editors (2003). *Mobility Data: Modeling, Management, and Understanding*. Cambridge University Press.

Seep, J. and Vahrenhold, J. (2019). Inferring semantically enriched representative trajectories. In *Int. Workshop on Computing with Multifaceted Movement Data*, MOVE'19. DOI: 10.1145/3356392.3365220.