Towards a Representativeness Measure for Summarized Trajectories with Multiple Aspects

Vanessa Lago Machado[1,2], Tarlis Tortelli Portela[3], Chiara Renso[4], Ronaldo dos Santos Mello[1]

[1]Universidade Federal de Santa Catarina (UFSC), INE, Florianópolis, SC – Brazil

[2]Instituto Federal Sul-Rio-Grandense (IFSUL), Passo Fundo, RS – Brazil.

[3]Instituto Federal do Paraná (IFPR), Palmas, SC – Brazil

[4]Consiglio Nazionale delle Ricerche (CNR), ISTI, Pisa, Italy

vanessalagomachado@gmail.com, tarlis@tarlis.com.br

chiara.renso@isti.cnr.it, r.mello@ufsc.br

Abstract. *Large trajectory datasets have led to the development of summarization methods. However, evaluating the efficacy of these techniques can be complex due to the lack of a suitable representativeness measure. In the context of multi-aspect trajectories, current summarization lacks evaluation methods. To address this, we introduce RMMAT, a novel representativeness measure that combines similarity metrics and covered information to offer adaptability to diverse data and analysis needs. Our innovation simplifies summarization technique evaluation and enables deeper insights from extensive trajectory data. Our evaluation of real-world trajectory data demonstrates RMMAT as a robust Representativeness Measure for Summarized Trajectories with Multiple Aspects.*

1. Introduction

In an era of vast trajectory data generated by individuals, vehicles, and objects, the need to distill valuable insights is paramount. 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 efficiently comprehend and leverage the underlying movement patterns. 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.

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.

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.

2. Fundamentals

Geolocation services have become crucial in modern technology, leveraging vast amounts of data from large-scale tracking to monitor the movement of objects. This data is increasingly harnessed for purposes such as analysis, mining, and decision-making [Renso et al. 2013, 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 Guting [Erwig et al. 1999]. This raw trajectory comprised two dimensions: spatial and temporal. Around 2007, the notion of a *semantic trajectory* emerged. Here, 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 [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, Machado et al. 2023].

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

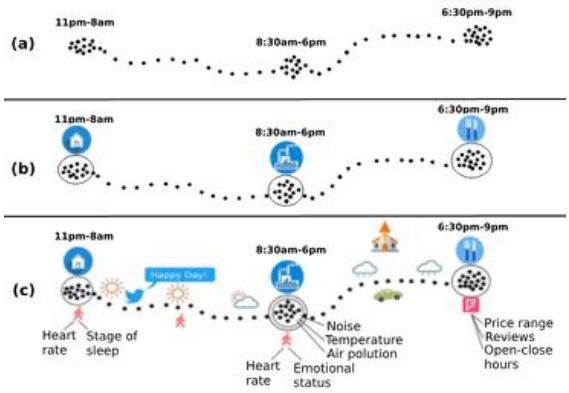

Figure 1. An example of a raw trajectory (a), semantic trajectory (b), and multiple aspect trajectory(c). Adapted from [Mello et al. 2019].

diversity of information [Martinez et al. 2018, Gao et al. 2019]. In this context, data summarization emerges as a viable strategy to condense similar trajectories and reduce the complexity of data management.

Trajectory summarization aims at reducing the volume of trajectory data while preserving its essential characteristics and patterns in a more compact representation [Hesabi et al. 2015]. Representative trajectories, in particular, provide a concise and informative presentation of a trajectory input dataset, facilitating analysis, visualization, and other trajectory-based tasks. In short, *MAT summarization* encompasses a process of abstraction from a set of MATs, culminating in a *representative MAT*. Notably, the representative MAT need not exhibit complete congruence with every individual MAT, but it captures the overarching essence of the dataset [Machado et al. 2022].

Understanding patterns in trajectories can help data analysts 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 offer a comprehensive insight 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.

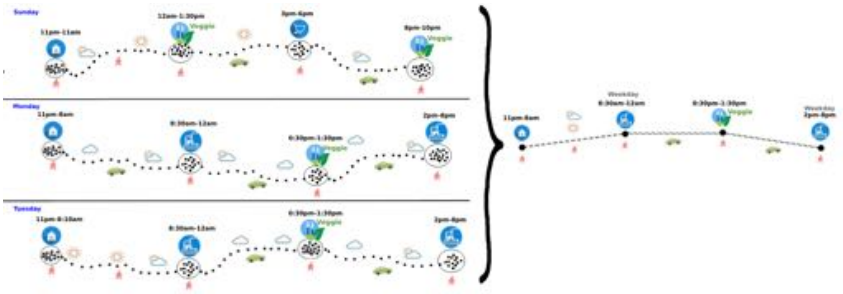

Figure 2. Examples of MATs (left) and a representative MAT for them (right) [Machado et al. 2022].

3. Problem Definition

In Figure 2, an example of trajectory summarization applied to input dataset D (D = $\{p, q, r\}$) generates the *Representative Trajectory (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 [Seep and Vahrenhold 2019, Machado et al. 2022] highlight this common challenge when computing representative trajectories from MATs.

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 *RTs* 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. We aim to facilitate the evaluation of summarization techniques and extract valuable insights from extensive MAT datasets.

4. RMMAT: Representativeness Measure for Multiple-Aspect Trajectory

In this section, we introduce the fundamental concepts of our work, which is called *RMMAT*[1]: a representativeness measure for MATs. We introduce a novel *Representativeness Measure* grounded in a similarity metric and covered information. By giving numerical values to the similarity, the measure provides a concrete and measurable way to measure how closely the *RT* reflects the complex patterns within the input dataset. By the covered information, this component enables us to examine whether the *RT* can encapsulate specific points from the input dataset, effectively reflecting the integrity of the *RT* concerning the entire dataset. By blending these two components, RMMAT aiming results in a rigorous and objective measure enables the evaluation of how well the *RT* captures the data's intricacies. This measure aims to overcome limitations in evaluating representativeness in summarized MAT.

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

[1]Source code available at `https://github.com/RepresentantativeMAT/RMMAT.git`

values or anomalies in similarity scores, resulting in a more balanced representation of central tendency. The equation is given by:

$$|Similarity\,(RT, D)| = Me(\{Similarity(RT, T_1), Similarity(RT, T_2), ..., Similarity(RT, T_n)\}) \quad (1)$$

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:

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

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 $\in$ [0,1]:

$$RMMAT = \omega_{sim} \times |Similarity\,(RT, D)| + \omega_{cover} \times \left(\frac{\sum_{p \in T}^{D} p \subseteq RT}{|D.points|}\right) \quad (3)$$

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

This section presents a running example of how RMMAT works and evaluates it through experimentation in a real dataset to assess its accuracy, robustness, and practicality in capturing trajectory data. The experiments were conducted on a Dell Inspiron laptop with an Intel Core i5 processor and 16 GB memory using 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 5.4 and 5.5) 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

In computing RMMAT, several key elements require definition: (i) the selection of a summarization method responsible for deriving representative data; (ii) the establishment of an appropriate similarity measure; (iii) the definition of weights (W) to individual components. We opt for the state-of-the-art MAT summarization method, *MAT-SGT* [Machado et al. 2023], and the widely recognized MAT similarity measure, *MUITAS* [Petry et al. 2019], to establish trajectory similarity. We employ a balanced weights strategy, setting $\omega_{sim} = \omega_{cover} = \frac{1}{2}$.

5.2.1. Summarization method setup

MAT-SGT summarizes 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.

We performed experiments by executing 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 for τ_{rv} and τ_{rc}, varying from 0% to 25% (0, 1, 5, 10, 15, 20, 25), to evaluate 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 as the baseline, computed across all trajectories in the group. This ensures that the baseline acts as a reference point for comparison purposes.

5.2.2. Similarity Measure setup

To compute similarity using MUITAS, 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. Since RT by MAT-SGT follows a different structure (rank values for categorical values of the semantic and temporal dimensions), analysis and different settings are required. Adopted functions are: (i) *Euclidean distance* for spatial dimension. A match occurs if the distance between a trajectory t_j in the group and RT coordinates is within a predefined threshold *(4 ×* $pointDispersionMeasure$*)*. The $pointDispersionMeasure$ is determined by the spatial dispersion of MAT points in 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.

5.3. Running Example

We introduce a Running Example to illustrate the functionality of RMMAT. It consists of a set of input MATs **D**, each representing a trajectory attributed to a different individual. The input MATs and their corresponding RT are shown in Figure 3. They are represented by spatial and temporal information, along with the price and category of the PoIs, weather conditions, and precipitation.

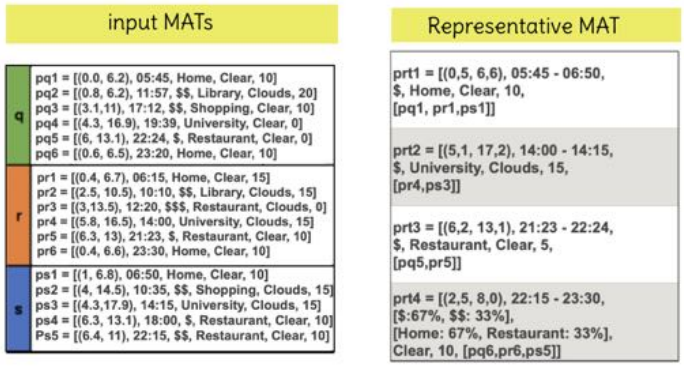

Figure 3. Set of input MATs D $= \langle q, r, s \rangle$**, where** $q = \langle p_{q_1}, p_{q_2}, ..., p_{q_n} \rangle$**,** $r = \langle p_{r_1}, p_{r_2}, ..., p_{r_m} \rangle$**, and** $s = \langle p_{s_1}, p_{s_2}, ..., p_{s_t} \rangle$ **(left), and their correspondent RT (right).**

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, $\left(\frac{\sum_{p \in T}^{D} p \subseteq RT}{|D.points|}\right) = \frac{10}{17} = 0.5882$.

Finally, considering the computation of RMMAT with balanced weights strategy by setting $\omega_{sim} = \omega_{cover} = \frac{1}{2}$ and according to Equation 3: $RMMAT = (0.5 \times 0.835) + (0.5 \times 0.5882) = 0.7116$, aiming that the RT have a representativeness measure of 0.7116 of D, considering both similarity and covered information.

5.4. Analyzing RMMAT Regarding Similarity Information

We analyzed a sample of user trajectories to gain insights into RMMAT behavior and 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, with a lower SD for average similarity scores; (ii) user 730, with a lower SD for median similarity scores; and (iii) user 708, 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 RC and RV, using $\omega_{sim} = 1$ and $\omega_{cover} = 0$ based on MUITAS. The investigation examines how different combinations of these thresholds affect the computation of RTs. Figure 4 shows the similarity evaluation results for each user with different input parameter configurations, compared to the baseline, while varying the temporal threshold. The threshold RC is abbreviated as tauRC.

Our RMMAT consistently outperformed the baseline for low parameter configurations. This analysis aims to provide insights into the interplay between different threshold

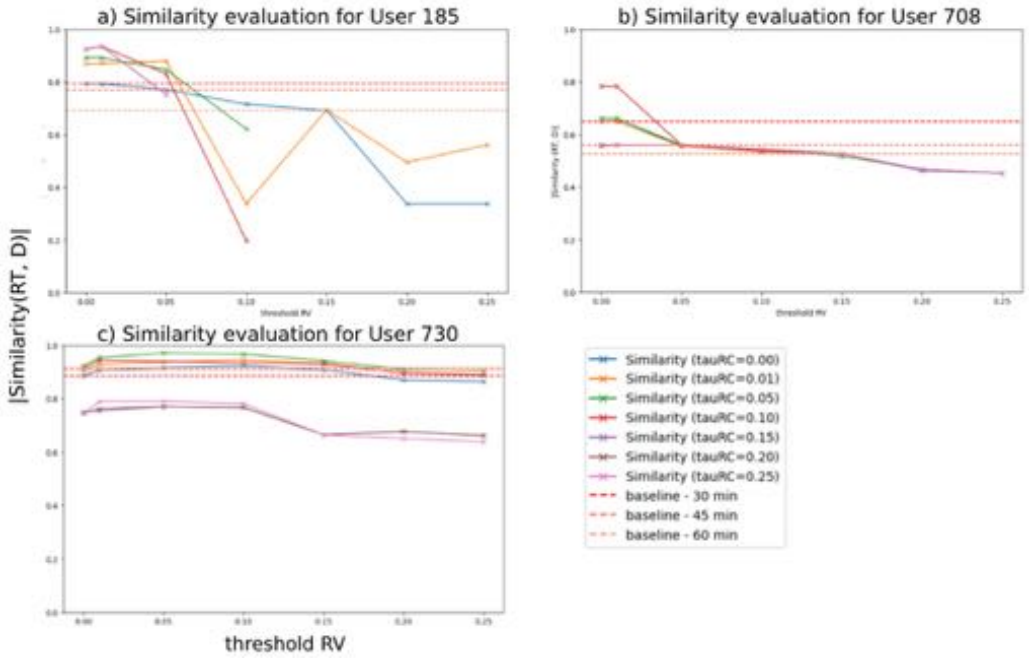

Figure 4. This graph analyzes the similarity evaluation (Y-axis) by comparing varying threshold RC (tauRC), shown as distinct lines, and the threshold RV in relation to baseline for users 185, 708, and 730. It explores different parameter configurations of the threshold RV (X-axis) to evaluate similarity.

parameters and their impact on RT computed from MUITAS. Users 708 and 730 exhibit a specific *RT* behavior pattern across different RV threshold values. Regarding the threshold RC, determining relevant cells for RT computation seems to influence RT changes significantly. As the value of this parameter configuration increases, RMMAT decreases. The behavior of user 185, in turn, underscores the impact of the choice of parameter configurations in RT computation concerning the representativeness of RT.

Using correlation coefficients, we analyzed how threshold values for RC and RV in MAT-SGT impact the RMMAT measure. These coefficients reveal relationships between input parameters and RMMAT scores for RT and input trajectories. Positive coefficients imply higher thresholds lead to higher RMMAT scores, while negative coefficients suggest the opposite. The results in Table 1 shed light on the influence of threshold parameters on the accuracy of computed representative trajectories.

Table 1. Impact of Input Parameters on the Representativeness Measure of RT

correlation coefficient	threshold RC	threshold RV
User 185	0.408	-0.788
User 708	-0.154	-0.829
User 730	-0.817	-0.243

User 185 exhibits a positive correlation (0.408) between RMMAT scores and threshold RC. The RMMAT scores increase as threshold RC values increase. User 708, characterized by greater SD in similarity scores, shows a slight negative correlation (-

0.154), indicating that increasing threshold 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 threshold RC values lead to lower RMMAT scores.

The threshold RC and RV significantly influence the behavior and accuracy of the computed representative trajectory. Understanding their impact helps make informed decisions about their selection to capture relevant input data patterns.

This analysis of RMMAT about similarity information provides valuable insights into the RT computation. It highlights the improvements achieved through the RMMAT measure and underscores its power in enhancing data comprehension. The results emphasize the effectiveness of RMMAT as a tool for gaining a deeper understanding of data.

5.5. Analyzing RMMAT Regarding Covered Information

To analyze the impact of covered information in RMMAT, we assess the utility of RT by employing the *Average Recall (AR)* metric in an experimental evaluation based on MUITAS. We adopted the MUITAS methodology and dataset for our evaluation. We intend to quantify the quality of RT summarization and representative data computation.

AR measures the recall based on the RMMAT computed between the *RT* and other MATs in the dataset. The objective is to ensure that the *RT* of each user achieves a high measure score when compared to MATs within the same group. This alignment stems from the likelihood that trajectories from the same user tend to exhibit higher scores.

To evaluate the recall information for each trajectory, we have modified an internal programming mechanism of MAT-SGT. This mechanism dynamically determines the optimal grid size for computing *RTs* by iteratively calculating it. Initially, this process only relied on the similarity measure. However, our modified approach now incorporates covered information in a balanced manner, taking advantage of the mapping data inherently present in MAT-SGT. This mechanism enables us to compute and evaluate this crucial aspect of representativeness comprehensively.

We tested two scenarios: (i) using the original MAT-SGT without covered information and (ii) using our adapted version of MAT-SGT with covered information. We evaluated the results by computing RT for each user group, calculating similarity using MUITAS, ordering trajectories based on similarity scores, and computing the recall metric. The recall metric measures the ability of RT to rank trajectories within the same group accurately.

Tables 2 and 3 show the AR values for user 185 in both scenarios, respectively. Table 4 compiles the results of the AR analysis. The variations are underlined in Tables 2 and 3. It is important to note that 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.

After analyzing the summarized outcomes of the AR analysis in Table 4, we observe some relevant variations between including and excluding covered information for User 185. Specifically, we see an average AR growth of 0.707 when analyzing the scenario without covered information, compared to 0.771 when combining covered information.

Table 2. The AR of User 185 - 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	-	-	-	-	-

Table 3. The AR of User 185 - 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	0.83	0	0.81	0	-	-	-
0.15	0.86	0.98	-	-	-	-	-
0.20	0.02	1	-	-	-	-	-
0.25	0.02	0.83	-	-	-	-	-

Table 4. AR Analysis regarding covered information in User 185

	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

In the case of User 708, Tables 5 and 6 present the AR values for both scenarios. Table 4 provides a summary of the AR analysis results for this user. Although some minor variations in specific values were observed, the overall assessment presented in Table 7 does not indicate a substantial difference. The AR values for this user remain relatively stable, irrespective of whether the covered information was included or excluded during the analysis.

Table 5. The AR of User 708 - 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

Table 6. The AR of User 708 - 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

Table 7. AR Analysis regarding covered information in User 708

	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

The AR values for User 730 in both scenarios are presented in Tables 8 and 9. Additionally, Table 10 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.

Table 8. The AR of User 730 - 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

Table 9. The AR of User 730 - 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

Table 10. AR Analysis regarding covered information in User 730

	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 inclusion or exclusion of covered point data presents a high impact for some users, like user 730, whose outcomes were notably affected. However, when considering covered point data, the retrieved trajectories from the same user exhibit better results than computed RT trajectories from the same user. It suggests that covered point data can affect RMMAT scores, indicating potential differences in underlying data patterns. This emphasizes the importance of considering each component in the RMMAT calculation to create a customized configuration that suits specific datasets and analysis objectives.

6. Conclusion

This paper introduces the RMMAT, a standardized metric for evaluating the effectiveness of representative data given by summarization methods. It measures 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.

RMMAT uses similarity metrics and covered information to provide a comprehensive evaluation approach. This helps analysts estimate the similarity between representative and input trajectories and the coverage of information within the dataset. This measure empowers researchers and analysts to make informed decisions regarding the quality and relevance of representative data for analytical goals.

RMMAT effectively quantifies the representativeness of computed representative data compared to the original MATs, yielding valuable insights. For instance, in the case of MAT-SGT, the evaluations highlighted the key role of parameter selection in achieving optimal results. This observation emphasizes how RMMAT offers insights that can guide researchers in refining their trajectory summarization methods for improved outcomes.

One of the notable strengths of RMMAT lies in its adaptability. The configurable nature of its components permits analysts to tailor the evaluation process to match the unique demands of different analytical scenarios, providing a versatile tool that aligns with varying objectives and data characteristics.

Our work bridges a critical gap in the field of trajectory data summarization, allow-

ing researchers and analysts to evaluate and measure trajectory summarization methods by a quantitative metric. By overcoming the limitations of previous subjective evaluation methods, RMMAT opens the door to more accurate and informed decision-making, deeper insights, and advancements in the field of data-driven mobility analysis.

The effectiveness of computing an *RT* depends on the specific use case, requiring varying levels of granularity and information preservation. The evaluation of this approach also depends on the purpose to be analyzed. This work focused on similarity and covered information, while future work aims to explore other views of summarized MAT representativeness, like reduced information.

Acknowledgments

This work was supported by CAPES - Finance Code 001, SoBigData++ Project - by TNA, and the EU's Horizon 2020 research and innovation programme under GA N. 777695 (EU Project MASTER). The views expressed in this paper are the authors' only responsibility.

References

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

Gao, C. et al. (2019). Semantic trajectory compression via multi-resolution synchronization-based clustering. *Knowledge-Based Systems*, 174:177–193.

Hesabi, Z. R. et al. (2015). Data summarization techniques for big data—a survey. In *Handbook on Data Centers*, pages 1109–1152. Springer.

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

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.

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. et al. (2019). MASTER: A multiple aspect view on trajectories. *Trans. GIS*, 23(4):805–822.

Oladimeji, D. et al. (2023). Smart transportation: an overview of technologies and applications. *Sensors*, 23(8):3880.

Parent, C. et al. (2013). Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4):42:1–42:32.

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

Renso, C., Spaccapietra, S., and Zimányi, E. (2013). *Mobility Data: Modeling, Management, and Understanding*. Cambridge University Press, Cambridge.

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