UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Vanessa Lago Machado

# On Computing Representative Data for Summarizing Multiple Aspect Trajectories

Florianópolis
2024

Vanessa Lago Machado

# On Computing Representative Data for Summarizing Multiple Aspect Trajectories

Tese submetida ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do título de Doutora em Ciência da Computação.
Orientador: Prof. Ronaldo dos Santos Mello, Dr.
Coorientadora: Prof[a]. Vania Bogorny, Dr[a].

Florianópolis
2024

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Machado, Vanessa Lago
On Computing Representative Data for Summarizing Multiple Aspect Trajectories / Vanessa Lago Machado ; orientador, Ronaldo dos Santos Mello, coorientador, Vânia Bogorny, 2024.
112 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2024.

Inclui referências.

1. Ciência da Computação. 2. Multiple-Aspect Trajectory. 3. Data Summarization. 4. Trajectory Summarization. 5. Representative Trajectory. I. Mello, Ronaldo dos Santos. II. Bogorny, Vânia. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. IV. Título.


Documento assinado digitalmente
**Vanessa Lago Machado**
Data: 02/07/2024 17:58:23-0300
CPF: ***.183.870-**
Verifique as assinaturas em https://v.ufsc.br

Vanessa Lago Machado

**On Computing Representative Data for Summarizing Multiple Aspect Trajectories**

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Profª. Carina Friedrich Dorneles, Drª.
Universidade Federal de Santa Catarina

Profª. Renata de Matos Galante, Drª.
Universidade Federal do Rio Grande do Sul

Prof. José Antonio Fernandes de Macêdo, Dr.
Universidade Federal do Ceará

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Doutora em Ciência da Computação.


Documento assinado digitalmente
**Marcio Bastos Castro**
Data: 03/07/2024 10:32:46-0300
CPF: ***.707.720-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Márcio Bastos Castro, Dr.
Coordenador do Programa


Documento assinado digitalmente
**RONALDO DOS SANTOS MELLO**
Data: 02/07/2024 19:16:47-0300
CPF: ***.693.660-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Ronaldo dos Santos Mello, Dr.
Orientador

Florianópolis, 2024.

I dedicate this thesis first and foremost to God, my loving and caring Father, and to Jesus Christ, my guide and model. I also extend my gratitude to my internal process of constant evolution and transformation, which has supported and inspired me throughout this academic journey.

## ACKNOWLEDGEMENTS

First and foremost, I extend my deepest gratitude to God, my loving and caring father, and Jesus Christ, my guide and model. I also wish to express my heartfelt thanks to my earthly parents, Sônia Mara Machado and Mário Dias Machado, for teaching and guiding me along this path.

I am immensely grateful to my advisor, Ronaldo dos Santos Mello, for his unwavering support, friendship, and patience throughout this journey. He has guided me in becoming a researcher and has been more than an advisor—he has been a true friend.

I also want to thank my co-advisor, Vânia Bogorny, who has inspired me as a researcher, provided valuable insights for my research and reminded me of what truly matters in life and what we should value. Similarly, I am grateful to my research supervisor in Italy, Chiara Renso, for her warm reception and guidance during the research conducted through the SoBigData project at CNR in Pisa. I also thank my colleagues at the CNR laboratory and researchers (especially Chiara, Guido, Nelson, Joaquin, Vinicius, Jonatan, and Raffaele) for welcoming me with open arms and making this research period more enjoyable.

I am grateful to Instituto Federal Sul-Rio-Grandense (IFSUL), where I work as a professor, for granting me the license to conduct this research. Without their support, this research would not have been possible. I would also like to thank the SoBigData++ Project - Transnational Access (TNA), which is part of the European Union's Horizon 2020 research and innovation programme, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), and the Universidade Federal de Santa Catarina (UFSC) for their belief in this research and for enabling it through the MASTER project.

I am deeply grateful to my colleagues in the UFSC laboratory, especially Ana Paula, Suzane, Geomar, Angelo, and Tarlis, for all their assistance during the laboratory exchanges, making this achievement more friendly through our interactions. Special thanks to my great friend Tarlis for all our exchanges, friendship, and study hours in coffee shops, which made the pandemic period more bearable and filled with insights. I also thank my friend Ana Caroline for our deep talks, exchanges, and study moments in coffee shops, showing me that everything is part of a greater process.

A heartfelt thank you to my cousin Gabriel de Andrade for a lifetime of brotherly bond and, especially during this period, for his unwavering support and closeness, reminding me that I can always count on him. Thank you for being a shoulder to lean on and always being there for conversations, coffee, wine, and deep talks. More than a cousin, you are a brother at heart.

Symbolically, I want to thank the city of Florianópolis, a magical island, for receiving me during this time. It has allowed me to grow in my research and personal journey, helped me in my self-knowledge, and for all the friends that have given me during this period. Additionally, I am deeply grateful to my pets, Spike (in memory), Vicky, and Layla, for the love they have shared with me throughout our lives together. I am deeply grateful to be their caretaker.

Once again, I want to express my gratitude to my family who are the reason for who I

am today. They have enabled me to dream and achieve my dreams. Some of them are no longer with us (my father, Mário, and Grandma Julieta), but their memory and teachings have shown me the importance of our family heritage. I thank them for everything; I am who I am because of them, especially my mother, Sônia, for believing that her children could dream and achieve their dreams, and my father, Mário, for all the love and protection he gave.

*Eu gostaria de expressar minha mais profunda gratidão à minha família, a razão de quem eu sou hoje. Eles me permitiram sonhar, almejar e alcançar esses sonhos. Algumas pessoas queridas já não estão mais entre nós, como meu pai, Mário, e minha avó, Julieta, mas suas memórias e ensinamentos me mostraram a importância da ancestralidade em cada um de nós. Sou imensamente grato a eles por tudo. Sou quem eu sou graças à minha família, especialmente à minha mãe, Sônia, por acreditar que seus filhos podiam sonhar e alcançar seus objetivos, e ao meu pai, Mário, por me ensinar o que é o amor e pela proteção que sempre me proporcionou.*

"If you cannot find the path, make one. Cut through. Go around. Tunnel under.
Fly above. There is always a way through if the way is true."
(Clarissa Pinkola Estés)

## RESUMO

Nos últimos anos, houve um aumento significativo na coleta de dados de mobilidade, impulsionado pela proliferação da Internet das Coisas. Esses dados abrangem uma ampla gama de fontes, fornecendo informações detalhadas sobre movimento e localização ao longo do tempo, formando o que é denominado como trajetória de objetos móveis. Esses dados de mobilidade não se limitam apenas à sequência de movimentação no espaço e tempo, mas também englobam uma variedade de aspectos relacionados ao objeto em movimento, ao ambiente e ao trajeto em si, originando as chamadas trajetórias de múltiplos aspectos. Por exemplo, ao rastrear a trajetória de movimento de um indivíduo ao longo do dia, é possível capturar não apenas informações sobre sua localização, mas também dados relacionados à sua saúde, condições climáticas, locais visitados e modos de transporte utilizados. Essa abundância de dados de mobilidade proporciona perspectivas promissoras para análises mais aprofundadas e compreensão das dinâmicas de movimento em diferentes domínios de aplicação, incluindo controle de tráfego, previsão de eventos extremos (como furacões e tsunamis), sistemas de recomendação, entre outros. No entanto, lidar eficientemente com esses vastos volumes de dados heterogêneos representa um desafio considerável, dificultando a extração de *insights* valiosos, tanto devido à complexidade dos dados quanto ao seu processamento. Nesse contexto, a sumarização de trajetórias gerando dados representativos emerge como uma potencial solução para minimizar esses desafios na manipulação de dados de trajetórias com múltiplos aspectos. Os métodos atuais de sumarização de dados de trajetórias frequentemente se concentram apenas nas dimensões espacial e temporal, ignorando os múltiplos aspectos semânticos dos dados. Neste contexto, este trabalho propõe o desenvolvimento de novos algoritmos para sumarizar dados de trajetórias de múltiplos aspectos. Duas abordagens são apresentadas: MAT-SG, baseada na densidade espacial dos dados, e MAT-SGT, baseada na densidade espacial e temporal. Resultados experimentais demonstraram a eficácia das abordagens propostas em diferentes conjuntos de dados, destacando sua capacidade de fornecer uma representação significativa das trajetórias de mobilidade. Além disso, uma medida de representatividade é introduzida neste trabalho para avaliar a qualidade dos dados representativos gerados.



**Palavras-Chave:** Trajetória de Múltiplo Aspecto. Sumarização de Dados. Sumarização de Trajetórias. Trajetória representativa.

## RESUMO ESTENDIDO

### Introdução

Nos últimos anos, a proliferação da Internet das Coisas tem impulsionado um aumento significativo na produção e coleta de dados da mobilidade de objetos, como pessoas, animais ou veículos. Esses dados, conhecidos como trajetória de objetos móveis, oferecem uma visão do movimento e posição ao longo do tempo. Com o avanço das pesquisas nesta área, foi reconhecido o potencial de enriquecer esses dados espaço-temporais com informações semânticas, resultando no conceito de trajetórias semânticas. Mais recentemente, observou-se o potencial de enriquecer um ou mais pontos das trajetórias com diversos aspectos semânticos, conhecido hoje como a noção de trajetórias de múltiplos aspectos (MELLO et al., 2019). A Figura 1 ilustra a evolução ao longo dos anos dessas categorizações associadas a trajetórias.

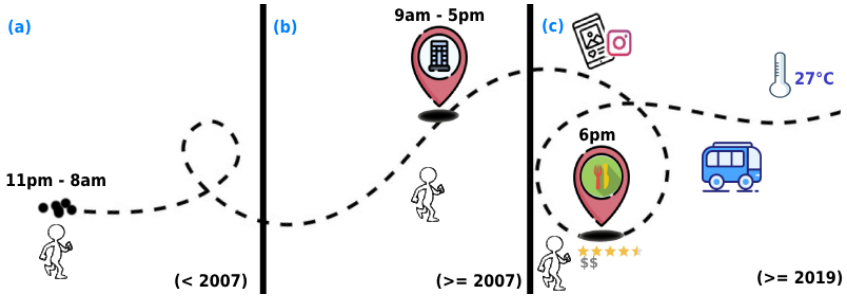


Figura 1 – Evolução histórica dos tipos de Trajetórias

A Figura 1(a) ilustra a trajetória de um indivíduo ao longo do dia por meio de uma trajetória denominada bruta, que inclui informações sobre sua mobilidade, sua posição geográfica e o tempo. Na Figura 1(b), é apresentada a mobilidade desse indivíduo por meio de uma trajetória semântica, enriquecida com informações sobre sua mobilidade espaço-temporal e os locais por ele visitados. Por fim, a Figura 1(c) demonstra a mobilidade desse indivíduo por meio de uma trajetória de múltiplos aspectos, que também incorpora informações sobre suas postagens em redes sociais, como avaliações dos locais visitados, condições climáticas e meio de transporte utilizado.

Essa vasta quantidade de dados gerados continuamente e a complexidade desses dados de múltiplos aspectos introduzem desafios na sua gerência e análise. Neste contexto, a sumarização de trajetórias surge como uma potencial solução para lidar com dados de trajetórias de múltiplos aspectos. A sumarização de dados de trajetórias pode ser definida como uma técnica para resumir os dados de trajetórias, com o objetivo de evidenciar informações mais relevantes e gerenciar melhor o volume de dados (ETIENNE et al., 2016). Esta técnica visa reduzir o volume de dados (FENG; ZHU, 2016) enquanto preserva os padrões principais da mobilidade original (AHMED, 2019). Um dado representativo é considerado aquele que captura o movimento principal de um conjunto de trajetórias (LEE; HAN; WHANG, 2007; AYHAN; SAMET, 2015). Portanto, a sumarização de dados com o intuito de computar uma informação representativa pode beneficiar diversas aplicações, como sistemas de recomendação, previsão de fenômenos naturais e detecção de anomalias.

O principal desafio na sumarização de dados de trajetórias de múltiplos aspectos está na complexidade desses dados, que envolve uma grande quantidade de informações e heterogeneidade nas dimensões associadas a cada ponto da trajetória. Por exemplo, um ponto de interesse (POI) pode agregar vários contextos semânticos do local visitado, como dimensões espaciais (latitude e longitude) e dados como categoria (um hotel, por exemplo), preço e avaliação do local. Além

disso, os pontos da trajetória podem conter informações sobre o indivíduo naquele momento, como batimentos cardíacos, e informações ambientais, como condição climática. Portanto, a sumarização desses dados apresenta desafios significativos.

**Objetivos**

O objetivo desta tese é desenvolver um *framework* composto por novos métodos para sumarizar dados de trajetórias de múltiplos aspectos, visando reduzir dados e capturar informações essenciais, computando um dado representativo para um conjunto de trajetórias similares.
Para atender a este objetivo principal foram propostos os seguintes objetivos específicos:

- Propor e implementar algoritmos de identificação de densidade de trajetórias de múltiplos aspectos para serem sumarizadas;
- Propor e implementar métodos para sumarizar todos os aspectos das trajetórias de entrada visando tratar a individualidade de cada aspecto, crucial para garantir que todos os aspectos das trajetórias sejam adequadamente considerados durante o processo de sumarização, permitindo uma representação mais completa e precisa dos dados;
- Propor uma nova medida de representatividade permitindo avaliar quantitativamente a qualidade de uma trajetória de múltiplos aspectos representativa.

**Metodologia**

A seguinte metodologia é adotada neste trabalho para alcançar os objetivos propostos:

1. Realizar revisão de literatura em sumarização de dados de trajetórias, com foco em trabalhos que resultam em dados representativos em dados de trajetórias multiplos-aspectos;
2. Propor um modelo conceitual para representação do dado representativo, visando manter o mapeamento entre as trajetórias originais e o dado representativo computado;
3. Propor e implementar algoritmos para identificação de densidade dos dados, focando na densidade espacial (MAT-SG) e densidade espaço-temporal (MAT-SGT);
4. Propor e implementar um novo método para sumarização de dados de trajetória de múltiplos aspectos baseado na densidade espacial, tratando os aspectos em sua individualidade;
5. Propor e implementar um novo método para sumarização de dados de trajetória de múltiplo-aspecto baseado na densidade espacial e temporal, tratando todos os aspectos em sua individualidade, resultando na sequência temporal do comportamento da trajetória;
6. Realizar uma série de experimentos utilizando conjuntos de dados de diferentes tipos e características.
7. Propor uma medida de representatividade visando mensurar a qualidade do dado representativo em função do conjunto de trajetórias originais, baseado na sua similaridade e informações cobertas pelo dado representativo;
8. Avaliar o comportamento dos métodos propostos usando conjuntos de dados reais e sintéticos, por meio de cálculos estatísticos. Avalia-se o desempenho dos métodos propostos em relação à capacidade de fornecer representações significativas das trajetórias;

9. Escrever artigos descrevendo as lacunas identificadas no estado-da-arte em relação a sumarização de trajetórias de múltiplos aspectos, bem como acerca dos novos métodos propostos visando computar a trajetória representativa;

10. Escrever a redação da tese descrevendo os principais conceitos necessários de dados de trajetória, o problema de sumarização, o estado da arte, a descrição das soluções propostas, avaliações experimentais e as conclusões obtidas.

Esta tese possui algumas limitações, as quais faz-se nessário estabelecê-las para uma melhor compreensão. Primeiro, seu foco é o desenvolvimento de novos métodos de sumarização de dados de trajetórias de múltiplos aspectos. Tal delimitação permite explorar a complexidade associada a estes dados. Segundo, utiliza-se o termo *redução de dados de trajetórias* nesta tese para se referir unicamente ao conceito de sumarização, compreendendo que a versão sumarizada dos dados minimiza o volume de dados. Terceiro, assume-se que os conjuntos de dados a serem sumarizados já se encontram filtrados por algum critério, exibindo assim algum grau de similaridade definida pelo analista. Desse modo, esta tese não lida com questões de limpeza de dados ou pré-processamento, concentrando-se nas atividades de sumarização.

**Resultados e Discussão**

Esta tese possui como principal contribuição um *framework* composto por dois novos métodos para sumarizar dados de trajetórias de múltiplos aspectos: MAT-SG e MAT-SGT, e uma medida de representatividade (RMMAT). Ambos os métodos são desenhados para prover dados representativos do conjunto original.

Em contraste com o estado-da-arte, que muitas vezes negligencia o tratamento dos múltiplos aspectos dos dados, ou mesmo suas particularidades, esses dois métodos visam abstrair cada uma das dimensões de acordo com sua própria singularidade, bem como capturar a sequência temporal dos dados. Além disso, os métodos propostos distinguem-se por manter um mapeamento claro entre os dados originais e os dados sumarizados, por meio de uma modelagem de fácil compreensão. Isso permite a persistência dos dados, facilitando a busca por padrões e *insights*. Ainda, uma vez que não encontramos na literatura uma forma de quantificar o quanto esse dado representativo reflete do conjunto original, como uma contribuição secundária esta tese também apresenta uma medida de representatividade (RMMAT) para avaliar a qualidade do dado representativo em relação à similaridade da trajetória representativa e sua cobertura de informação em relação ao conjunto original de trajetórias.

Experimentos foram conduzidos em quatro conjuntos de dados, incluindo conjuntos de dados abertos (*Foursquare-NYC, Gowalla Location-Based Social Network* e *Brightkite*) e um conjunto de dados privado (*dataset Pisa*). Todos os conjuntos de dados consistiam em trajetórias de usuários, e em todos os casos, foram filtrados os dados por usuário, calculando a trajetória representativa por usuário. Para avaliar a eficácia dos métodos propostos, foram utilizadas duas métricas: *(i) Average Recall* (AR), que avalia a capacidade dos métodos em ranquear corretamente as trajetórias do mesmo usuário como mais similares à trajetória representativa; e *(ii)* a métrica *RMMAT* para avaliar a qualidade do dado representativo em relação à similaridade e cobertura de informação das trajetórias do mesmo usuário.

Os resultados experimentais revelaram a eficácia das abordagens propostas em diversos cenários de dados. Tanto o método baseado em densidade espacial (MAT-SG) quanto o método que considera densidade espacial e temporal (MAT-SGT) demonstraram ser capazes de fornecer representações significativas das trajetórias com múltiplos aspectos. Além disso, a introdução da medida de representatividade permitiu uma avaliação mais precisa da qualidade dos dados representativos gerados.

Para superar algumas limitações identificadas nesta pesquisa, sugerem-se trabalhos futuros: *(i)* o desenvolvimento de estratégias de sumarização que considerem possíveis dependências entre diferentes aspectos dos dados, como a avaliação ou preço de um local visitado em detrimento ao local visitado em si; e *(ii)* a investigação de novas estratégias de segmentação espacial para reduzir a complexidade dos métodos.

**Considerações Finais**

Este trabalho contribui significativamente para o campo da sumarização de dados de trajetórias de múltiplos aspectos, fornecendo um *framework* composto por novos métodos para lidar com dados de mobilidade complexos. Os resultados obtidos sugerem que os métodos propostos são promissores e podem ser aplicados em uma variedade de domínios de aplicação.

**Palavras-Chave:** Trajetória de Múltiplo Aspecto. Sumarização de Dados. Sumarização de Trajetórias. Trajetória representativa.

## ABSTRACT

In recent years, the widespread adoption of the Internet of Things has led to a significant increase in the production and collection of mobility data. Various sources have provided this data, which provides comprehensive details about data movement and position over time, commonly referred to as the trajectory of moving objects Mobility data not only encompasses space and time but also includes multiple aspects related to the movement object, the environment, and the trajectory, resulting in multiple-aspect trajectories. For instance, by analyzing the trajectory movement of an individual during one day, it is possible to identify information about her/his position, time occurrence, health, weather conditions, visited places, and transportation modes. This large volume of data provides diverse perspectives for analyzing and understanding movement dynamics across various application domains, such as traffic control, forecasting extreme events (such as hurricanes and tsunamis), recommendation systems, and more. However, managing trajectory data poses challenges, making it difficult to efficiently extract valuable insights due to data complexity and processing requirements. In this context, trajectory summarization, which computes representative data, emerges as a potential solution to mitigate these challenges in handling multiple-aspect trajectory data. State-of-the-art methods often focus only on spatial and temporal dimensions, overlooking multiple semantic aspects. Hence, the objective of this thesis is to develop new algorithms for summarizing multiple-aspect trajectories by computing representative data. Our main contributions involve two novel methods: MAT-SG, based on spatial density, and MAT-SGT, based on both spatial and temporal density. Experimental results have demonstrated the efficacy of both proposed methods across different dataset types, highlighting their ability to provide a significant representation of input data. Additionally, a representative measure is introduced to evaluate the quality of computed data representatives.



**Keywords:** Multiple-Aspect Trajectory. Data Summarization. Trajectory Summarization. Representative Trajectory.

## LIST OF FIGURES


Figura 1 – Evolução histórica dos tipos de Trajetórias . . . . . 13
Figure 2 – Generic Process for Data Summarization . . . . . 37
Figure 3 – An example of a raw trajectory (a), semantic trajectory (b), and a multiple aspect trajectory(c). . . . . 39
Figure 4 – An example of MATs (a), a representative MAT for them (b), and an example of recommendation based on its representative MAT (c). Adapted from Machado, Mello e Bogorny (2022a). . . . . 41
Figure 5 – The conceptual model for MAT-SG. . . . . 54
Figure 6 – MAT-SG overview. . . . . 55
Figure 7 – Cell size computation . . . . . 58
Figure 8 – An example of temporal dimension summarization in a grid cell . . . . . 60
Figure 9 – The conceptual model for MAT-SGT . . . . . 64
Figure 10 – Overview of the MAT-SGT method. . . . . 65
Figure 11 – Sample data with point aspects information for trajectories $q$, $r$, and $s$. . . . 69
Figure 12 – Visualization of the resulting MAT-SG representative trajectory ($RT$) from different perspectives: (a) Spatial view; and (b) Detailed $RT$ description of point aspects, providing additional insights. . . . . 70
Figure 13 – A step-by-step perspective of the summarization process in MAT-SG, illustrated by the analyzed cell (a), the $p_r$ Computation step (b), and the final representative points computed (c). . . . . 70
Figure 14 – Resulting MAT-SGT in representative trajectory ($RT$) visualization in different perspectives: (a) Spatial perspective; (b) Spatiotemporal perspective; and (c) $RT$ description of point aspects providing additional details. . . . . 71
Figure 15 – A step-by-step perspective of the summarization process in MAT-SGT, illustrated by the analyzed cell (a), the Temporal Definition step (b), the $p_r$ Computation step (c), and the final representative points computed (d). . . . 72
Figure 16 – 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 $s = \langle p_{s_1}, p_{s_2}, ..., p_{s_t} \rangle$ (left), and their correspondent $RT$ (right). . . . . 77
Figure 17 – This graph analyzes the similarity evaluation (Y-axis) by comparing varying threshold RC, the $\tau_{rc}$, shown as distinct lines, and the threshold RV, the $\tau_{rv}$, concerning baseline for users 185, 708, and 730. It explores different parameter configurations of the $\tau_{rv}$ (X-axis) to evaluate similarity. This analysis refers to the MAT-SG method. . . . . 79
Figure 18 – This graph analyzes the similarity evaluation (Y-axis) by comparing varying threshold RC, the $\tau_{rc}$, shown as distinct lines, and the threshold RV, the $\tau_{rv}$, concerning baseline for users 185, 708, and 730. It explores different parameter configurations of the $\tau_{rv}$ (X-axis) to evaluate similarity. This analysis refers to the MAT-SGT method. . . . . 80

## LIST OF ALGORITHMS

Algoritmo 1 – MAT-SG . . . . . 57
Algoritmo 2 – MAT-SG:*cellGridAllocation* . . . . . 58
Algoritmo 3 – MAT-SG:*computeTemporalDimension* . . . . . 60
Algoritmo 4 – MAT-SGT . . . . . 66

## LIST OF SYMBOLS

| | |
|---|---|
| AR | Average Recall |
| MAT-SG | Multiple Aspect Trajectory Summarization based on a spatial Grid |
| MAT-SGT | Multiple Aspect Trajectory Summarization based on a spatial Grid and Temporal sequence |
| $RT$ | Representative Trajectory |
| RC | relevant cell |
| RV | representativeness value |
| $p_r$ | representative point |
| RMMAT | Representativeness Measure for Multiple-Aspect Trajectory |
| $sti$ | each significant temporal interval |
| $STI$ | a set of $sti$ (significant temporal intervals) |
| $\mathbf{T}$ | set of filtered MATs |
| $ts$ | a timestamp in a trajectory point |
| $\|T.points\|$ | size of all input MAT points |
| $T^c(RT)$ | covered information of $\mathbf{T}$ by $RT$ |
| $V_{\Delta_{Time}}$ | valid time interval set |
| $\delta_i$ | time difference between two consecutive timestamps $ts$s |
| $\Delta_{Time}$ | a set of $\delta$ values |
| $\tau_{rc}$ | Minimum proportion of $\|T.points\|$, deciding if a cell is considered a relevant cell to compute $p_r$ |
| $\tau_{rv}$ | A rate of representativeness value for ranking values by data frequency for summarization step |
| $\tau_s$ | minimum spatial threshold |
| $\tau_t$ | threshold of temporal time - used to define when the difference of two $ts$s is considered an $sti$ |

**CONTENTS**


**1 INTRODUCTION . . . . . 29**
1.1 PROBLEM STATEMENT . . . . . 30
1.2 OBJECTIVES . . . . . 31
1.3 CONTRIBUTIONS . . . . . 32
1.4 SCOPE DELIMITATION . . . . . 32
1.5 THESIS STRUCTURE . . . . . 33

**2 BASIC CONCEPTS . . . . . 35**
2.1 DATA SUMMARIZATION . . . . . 35
**2.1.1 Data Summarization vs. Data Compression . . . . . 35**
**2.1.2 Data Summarization vs. Data Fusion . . . . . 36**
**2.1.3 Classifying Data Summarization . . . . . 36**
**2.1.4 The Effectiveness of Data Summarization . . . . . 37**
**2.1.5 The Data Summarization Process . . . . . 37**
2.2 TRAJECTORY DATA . . . . . 38
**2.2.1 Applications and Challenges . . . . . 39**
2.3 TRAJECTORY SUMMARIZATION . . . . . 39
**2.3.1 Representative Trajectory Data . . . . . 40**
**2.3.2 Similarity Measures . . . . . 41**

**3 RELATED WORK . . . . . 43**
3.1 SURVEYS ON TRAJECTORY DATA . . . . . 43
3.2 RELATED WORKS ON TRAJECTORY SUMMARIZATION . . . . . 45
**3.2.1 Discussion . . . . . 49**
**3.2.2 Evaluation of the Representative Trajectory Data Computation Process 50**
**3.2.3 Summary . . . . . 52**

**4 METHODS FOR MULTIPLE ASPECT TRAJECTORY DATA SUMMARIZATION . . . . . 53**
4.1 MAT-SG: MULTIPLE ASPECT TRAJECTORY SUMMARIZATION BASED ON A SPATIAL GRID . . . . . 53
**4.1.1 Data model . . . . . 54**
**4.1.2 Architecture . . . . . 55**
**4.1.3 Algorithm . . . . . 56**
*4.1.3.1 Data Segmentation Component . . . . . 57*
*4.1.3.2 Representative Point Computation component . . . . . 59*
*4.1.3.3 Computation of the Better Representative Trajectory . . . . . 62*

4.2 MAT-SGT: MULTIPLE ASPECT TRAJECTORY SUMMARIZATION BASED ON A SPATIAL GRID AND TEMPORAL SEQUENCE . . . . . 62
**4.2.1 Data model . . . . . 63**
**4.2.2 Architecture . . . . . 64**
**4.2.3 Algorithm . . . . . 65**
*4.2.3.1 Data Segmentation Component . . . . . 66*
*4.2.3.2 Representative Point Computation Component . . . . . 67*
*4.2.3.3 Computation of the Better Representative Trajectory . . . . . 67*
4.3 OUTPUT DATA . . . . . 68
4.4 RUNNING EXAMPLE . . . . . 69
**4.4.1 MAT-SG . . . . . 69**
**4.4.2 MAT-SGT . . . . . 70**
4.5 SUMMARY . . . . . 71

**5 RMMAT: REPRESENTATIVE MEASURE FOR MULTIPLE ASPECT TRAJECTORIES . . . . . 75**
5.1 SIMILARITY METRIC COMPONENT . . . . . 75
5.2 COVERED INFORMATION COMPONENT . . . . . 76
5.3 RUNNING EXAMPLE . . . . . 77
5.4 ANALYZING RMMAT REGARDING SIMILARITY INFORMATION . . 78
5.5 ANALYZING RMMAT REGARDING COVERED INFORMATION . . . . 81
5.6 PROPERTIES OF RMMAT . . . . . 86

**6 EXPERIMENTAL EVALUATION . . . . . 87**
6.1 DATASETS . . . . . 87
6.2 METHODOLOGY . . . . . 88
**6.2.1 Evaluation Metrics . . . . . 88**
**6.2.2 Experimental Setup . . . . . 90**
6.3 RESULTS . . . . . 91
**6.3.1 AR Metric Strategy . . . . . 91**
*6.3.1.1 Foursquare-NYC dataset . . . . . 91*
*6.3.1.2 Gowalla Location-Based Social Network dataset . . . . . 91*
*6.3.1.3 Brightkite dataset . . . . . 92*
*6.3.1.4 Pisa dataset . . . . . 92*
**6.3.2 RMMAT Strategy . . . . . 93**
*6.3.2.1 Foursquare-NYC dataset . . . . . 94*
*6.3.2.2 Gowalla Location-Based Social Network dataset . . . . . 95*
*6.3.2.3 Brightkite dataset . . . . . 95*
*6.3.2.4 Pisa dataset . . . . . 96*
6.4 DISCUSSION . . . . . 98

**6.4.1 Limitations . . . . . 100**

**7 CONCLUSION . . . . . 103**

7.1 PUBLICATIONS . . . . . 105

**BIBLIOGRAPHY . . . . . 107**

# 1 INTRODUCTION

The rapid proliferation of the Internet of Things (IoT) has given rise to diverse technologies, including portable and wearable devices, embedded computing, and Location-Based Social Networks (LBSNs) like Facebook, Twitter, and Instagram (MUZAMMAL et al., 2017; CESARIO; COMITO; TALIA, 2014). These technologies yield valuable information on moving objects, such as people or animals. The collection of spatial position sequences over time forms the basis of a *raw trajectory* (BOGORNY; HEUSER; ALVARES, 2010). In the evolving geography of trajectory data, we recently encountered the concept of *Multiple Aspect Trajectory data (MAT)*, where trajectories encapsulate additional aspects such as visited places, health conditions, transportation modes, and weather conditions (MELLO et al., 2019). The accumulation of such data resulting from the movement of numerous objects can generate massive volumes of data.

Trajectory data has emerged as a focal issue in diverse domains, including data management (RICHLY, 2018; SU et al., 2020; WANG et al., 2021), data mining (FENG; ZHU, 2016; GEORGIOU et al., 2018; BIAN et al., 2018; da SILVA; PETRY; BOGORNY, 2019), privacy (FIORE et al., 2020), and monitoring (AHMED et al., 2019). Regarding trajectory management, the challenges are primarily associated with the large volume of continuously generated data and their diverse nature deriving from different devices and sources. Effective management and analysis of these data are critical for extracting valuable insights. Another challenge of complexity is related to the three dimensions inherent to MAT data (*spatial*, *temporal*, and *semantic*), where the third dimension is composed of multiple and heterogeneous aspects. In this context, *trajectory summarization* emerges as a potential solution to mitigate the complexity of manipulating MAT.

Trajectory summarization provides a promising route to address data management challenges, facilitating the extraction of meaningful patterns with applications across various domains. For instance, understanding individual behavior through trajectory data aids recommendation systems in delivering personalized suggestions. Furthermore, discerning patterns in weather conditions contribute to predicting and assessing the intensity of phenomena such as hurricanes.

The summarization of trajectory data poses a recognized challenge, as emphasized in various surveys (WANG et al., 2021; GEORGIOU et al., 2018; FIORE et al., 2020). However, there is a notable scarcity of literature that comprehensively presents and analyzes studies on trajectory summarization, particularly those that provide representative data, with a specific focus on MAT.

A *representative trajectory* refers to a compact yet informative representation of a set of trajectories, given typical patterns that capture the essential characteristics of the original dataset while minimizing information loss. For instance, the analysis of an individual's behavior is relevant to several application domains such as LBS recommendations and criminal investigations (FENG; ZHU, 2016). Despite the importance of this concept, there is a noticeable gap

in the literature, with only one identified summarization approach that provides representative data focusing on MAT (SEEP; VAHRENHOLD, 2019).

## 1.1 PROBLEM STATEMENT

For accurate trajectory summarization in high-dimensional trajectory datasets, the main challenge is to discover, in a feasible way, the most representative data considering their dimensions and aspects that better characterize the input data. By doing so, we can ensure that the trajectory data is summarized effectively.

Given a dataset **T** of trajectories, the problem of compute representative data (*RT*) can be formulated as follows

**Problem 1** (*Representative Trajectory Computation*)**.** A representative trajectory *RT* is a compact and informative representation of **T** that aims to strike a balance between quality and utility, ensuring that *RT* retains enough information about the original elements while minimizing data loss.

Let **T** be a set of trajectories. Given the large size of **T**, it is often necessary to summarize the trajectories to obtain a compact and informative representation of the original data, which makes less complex analysis or decision-making tasks. The problem of computing representative data can be formulated as the task of computing representative information from **T** that captures the essential characteristics of the original data while minimizing information loss. It is important to note that the concept of representativeness and capturing essential characteristics can be broad and generic. In this thesis, we specifically consider essential characteristics to be present in trajectories that exhibit a certain data density and show certain tendencies in their aspects.

Despite the importance of this issue, there is a noticeable gap in the literature, with only one summarization method providing representative data focused on MAT (SEEP; VAHRENHOLD, 2019). However, it is limited in the sense that all attributes of the points are treated as spatial or non-spatial data, i.e., semantic data are not analyzed individually as categorical or numeric data. It also does not provide details about the proposed method, as it is a short paper. In order to better address this problem, this thesis reviews and categorizes relevant research, aiming to provide a comprehensive understanding of trajectory summarization methods that yield representative data. This problem leads to our research question: **"Can we develop new methods for computing representative data for a set of MATs to discover relevant information and deal with gaps in related work by considering all aspects in MATs regarding their individuality?"**. We hypothesize that we can compute representative MAT by identifying patterns regarding some data density, summarizing all aspects considering their individuality, and providing utility data.

In order to tackle this question, we propose a *framework* composed by MAT-SG and MAT-SGT as novel trajectory summarization methods. MAT-SG is designed to address the

challenges associated with MATs by segmenting trajectories into a spatial grid and performing summarization within each relevant cell. This method aims to identify movement patterns specific to each spatial area, addressing multiple aspects and treating each one individually.

In contrast, MAT-SGT expands upon the methodology of MAT-SG by incorporating temporal sequence information into the summarization process. This enhancement allows MAT-SGT to provide a more comprehensive representation of the temporal evolution of movement patterns, thus capturing additional nuances in the data that may be overlooked by spatial-only data density.

The choice of summarization method depends on the intended use case. When prioritizing spatial areas and understanding the actions occurring in specific regions, such as in vessel trajectories, where it is essential to identify regions related to specific activities, like fishing or cargo handling, MAT-SG is the preferred method. On the other hand, when temporal sequence and the associated aspects are the focus, such as in recommendation systems where individual trajectories reveal patterns like daily routines and preferences based on weather conditions, MAT-SGT offers a more suitable solution.

Moreover, introducing these summarization methods prompted a secondary research question: **"How much of the representative trajectory captures and reflects the original MATs' essence within an input dataset?"**. This question underscores the need for a representativeness measure (RMMAT), motivated by the lack of quantitative measures comprehensively evaluating the quality of representative trajectory data. With RMMAT we aim to fill this gap by providing a multifaceted measure that assesses both the similarity and coverage of the representative trajectory in relation to the complete input dataset.

## 1.2 OBJECTIVES

The main objective of this thesis is to propose a *framework* composed of new methods for MAT summarization that address the gaps in the state-of-the-art, considering all aspects of MAT regarding their individually, while maintaining a focus on reducing data and capturing essential information from the input data. This thesis aims to contribute to the problem of summarizing MATs, considering that the concept of MAT and their data management is a brand new research topic. The main objective of this thesis is to propose *a framework composed by pioneering methods for MAT summarization that compute a representative MAT from a set of similar MATs*.

From this main objective, we can derive the following specific objectives:

- Propose and implement an algorithm for identifying the density of MATs to be summarized;
- Propose and implement methods for summarizing all aspects of the input MATs aiming to deal with their individuality;

- Propose a new representativeness measure to evaluate quantitatively the quality of representative MAT.

## 1.3 CONTRIBUTIONS

This thesis aims to contribute with state-of-the-art as follows:

- A survey related to state-of-the-art summarization of trajectory data focused on representative data computation;
- A conceptual model to represent the representative MAT, in the sense of defining mapping data between the input MATs and the representative MAT;
- A novel method for summarizing MAT data based on spatial density, treating all aspect data in its individuality;
- A novel method for summarizing MAT data based on spatial and temporal density, treating all aspect data in its individuality, providing the temporal sequence of the pattern.
- A multifaceted measure that assesses the quality of representative trajectory based on its representativeness (similarity and coverage information) of the complete input dataset.

The research results yielded by this thesis are intended to assist researchers and analysts with different approaches related to the use of MATs. It empowers them to make informed decisions about the quality and relevance of their data concerning the methods for summarizing MATs according to their analytical goals. Additionally, it provides a powerful tool with a measure to make informed decisions regarding the quality and relevance of representative data for analytical goals. With these contributions, researchers and analysts can analyze their data and compute and use the representative data in other approaches, such as performing predictions. It can also help in analyzing different sets of MATs and identifying their similarities, as well as analyze the quality and relevance of the data, empowering them to make informed decisions and achieve their analytical goals.

## 1.4 SCOPE DELIMITATION

To ensure a thorough understanding of this work, it is essential to establish some delimitations. These delimitations are crucial for the success of this research.

Firstly, our primary focus is on trajectories with multiple aspects. By considering various and distinct aspects enriched in trajectory data, the goal is to provide a more comprehensive analysis. This delimitation allows for a targeted exploration of the complexities associated with multiple-aspect trajectories.

Secondly, the term *reducing trajectory data* in this work refers only to the concept of trajectory summarization approaches. We understand that this strategy minimizes the volume of data, and our research delves into methodologies aimed at summarizing trajectories.

Thirdly, we assume that some criterion already filters the input trajectories and exhibits a degree of similarity. This assumption streamlines the focus on the analysis aspect, emphasizing the exploration of summarization techniques without delving into data-cleaning processes.

Finally, we also assume that the input data is already pre-processed. This pre-processing step ensures that the data is in a format ready for analysis. By making this assumption, the research can concentrate on the core aspects of trajectory summarization without being encumbered by data formatting concerns.

The research aims to provide a more focused and detailed investigation by establishing these delimitations. This focused strategy is anticipated to produce more accurate and reliable results, contributing to a fine understanding of multiple-aspect trajectory summarization.

## 1.5 THESIS STRUCTURE

The rest of this thesis is structured as follows. In Chapter 2, we discuss the primary concepts that help to understand our work. These concepts include data summarization, trajectory data, and trajectory summarization. We also analyze surveys on trajectory data to identify gaps in the literature. Next, in Chapter 3, we present the main works related to trajectory summarization to provide representative data.

In Chapter 4, we introduce two new methods for summarizing MATs. The first one is called MAT summarization based on a spatial grid (MAT-SG), which segments the input MATs into a spatial grid and performs summarization within each relevant cell. This helps to identify movement patterns specific to each spatial area, addresses various dimensions, and treats each semantic type individually. The second one is called MAT summarization based on a spatial grid and Temporal Sequence (MAT-SGT), which is a data summarization method specifically designed to compute representative MATs by identifying the temporal sequence associated with the movement pattern. We provide a running example to illustrate both methods and highlight their differences.

In Chapter 6, we present preliminary experiments. First, we introduce a multifaceted measure, the Representativeness Measure for MAT (RMMAT), that assesses the quality of a representative trajectory based on its representativeness (similarity and coverage information) of the complete input dataset. Then, we evaluate the experimental evaluation in several trajectory datasets using two different strategies: by Average Recall and by RMMAT. We then demonstrate that MAT-SG and MAT-SGT achieve good results in different dataset types.

Finally, in Chapter 7, we summarize the findings of this thesis and discuss future research opportunities in trajectory summarization that result in representative data.

# 2 BASIC CONCEPTS

In order to clarify the problem of trajectory summarization, this chapter presents the necessary concepts to understand the rest of this work. We start with an overview of *data summarization* and the elements that compose a data summarization process. Next, we introduce *trajectories of moving objects*, including MATs and some issues related to their summarization.

## 2.1 DATA SUMMARIZATION

*Data summarization* aims to provide data in a compact format, furnishing an informative version of a set of data. Providing a data summary is considered a descriptive task in data mining. A key feature of summarization is that this summarized representation of data is still informative, and a close inference (or sometimes the same inference) can be obtained from the summarized data in the same way as the original data (HESABI et al., 2015).

Consider $E = \{e_1,e_2,...,e_n\}$ a set $E$ of $n$ elements. Then, data summarization is formally defined as follows.

**Definition 2.1.1** (*Data Summarization*)**.** A summary $S$ of $E$ is a set of summarized elements $S = \{s_1,s_2,..,s_k\}$, where: $(i)$ $S$ is a non-empty set, $(ii)$ each $s_i \in S$ represents a summarized element of E, $(iii)$ each element $s_i \in S$ corresponds to at least one element $e_j \in E$, and $(iv)$ $k \leq n$.

A good summary is characterized by a small size for the summarized version while retaining enough information about all original elements. Each summary $s_i$ essentially covers a set of elements with minimal information loss. In other words, the elements of $E$ are summarized in a way that these elements are replaced by the corresponding summary that covers them (CHANDOLA; KUMAR, 2007). A summarized element can be derived from either a selected element $e_j \in E$ or through the computation of an element group, typically facilitated by statistical functions such as maximum or average.

### 2.1.1 Data Summarization vs. Data Compression

It is crucial to distinguish data summarization from data compression. *Data compression* is used to reduce data volume, where they consider compression techniques using statistical or dictionary-based methods, and they treat data as large byte sequences (AHMED, 2019). The formal definition of *Data Compression* is as follows:

**Definition 2.1.2** (*Data Compression*)**.** A compressed data $C$ of $E$ is given by $C = (C_E, decoder)$, where (i) $C_E$ is an encoded version of $E$, (ii) $C_E$ is a representation of $E$ with fewer bits, and (iii) $decoder$ is an algorithm that reconstructs $C_E$ in $E$ or some approximation of it.

Data compression entails encoding the original data, which converts the original data into a compressed representation, and decoding it to recover the information that reconstructs

the original data or an approximation from the compressed representation (BLELLOCH, 2013). Although compression produces compact data, it often results in compact yet unintelligible data. In contrast, data summarization offers an intelligible representation, facilitating further analysis and decision-making (AHMED, 2019).

### 2.1.2 Data Summarization vs. Data Fusion

Another clarification point is the similarities and differences between data summarization and data fusion. Data fusion integrates data from multiple sources to enhance accuracy and specificity compared to a single source (ESTEBAN et al., 2005). The formal definition is:

**Definition 2.1.3** (Data Fusion)**.** Given two sets $A$ and $B$, the fused data of $A$ and $B$ is a set $F_{(A,B)} = \{f_1, f_2, .., f_k\}$, where: $(i)$ $F_{(A,B)}$ is a non-empty set, and $(ii)$ each $f_i \in F_{(A,B)}$ represents an element matching pair $(a,b)$.

Effective data fusion results in a smaller $F_{(A,B)}$ size than $A \cup B$ while preserving information. In simpler terms, it means that when we combine two datasets, we can get a smaller dataset with minimal information loss. This means that the original elements of $A$ and $B$ are fused so that they are replaced by the fused element that covers them.

In summary, while both *data summarization* and *data fusion* produce representative versions of datasets, they differ in terms of the nature of the input data. Data fusion involves integrating data from multiple sources with similar information, while data summarization focuses on condensing a single dataset. In the context of this thesis, the emphasis is on data summarization.

### 2.1.3 Classifying Data Summarization

Ahmed (2019) classifies data summarization techniques into two categories: *structured* and *unstructured* data. Structured data refers to predefined formats featuring fixed fields or attributes with well-defined data types and relationships, typically following a predefined schema or model. Within structured data, subcategories include machine learning, statistical, and semantics approaches. Unstructured data, however, lacks a predefined data model or organized format. Examples encompass text documents, emails, social media posts, images, videos, and audio files. Subcategories within unstructured data cover machine learning and other diverse approaches.

Data summarization techniques can be categorized into two primary approaches based on their output: *extractive* and *abstraction* (GHODRATNAMA et al., 2020; MOHSIN et al., 2021). Extractive summarization entails selecting and presenting only the most pertinent information from the source data, efficiently eliminating redundancy. The goal here is to preserve the original data faithfully. Abstractive summarization, conversely, involves a deeper understanding of the meaning of data sources and generates new information that captures critical insights.

The emphasis is on creating a concise and coherent summary, even if it does not replicate the exact data or order from the source.

### 2.1.4 The Effectiveness of Data Summarization

The effectiveness of summarization data depends on the purpose for which they are used (GHODRATNAMA et al., 2020). Whether summarizing text (HEU; QASIM; LEE, 2015; MOHSIN et al., 2021; MA et al., 2022), documents (BOUDIN; HUET; TORRES-MORENO, 2011; GHODRATNAMA et al., 2020), images (SREELAKSHMI; MANMADHAN, 2021), or other data types, the chosen approach should align with the desired outcomes. For instance, a text summary helps readers learn essential points within a vast text, while a network traffic summary aids network administrators in understanding network activities (AHMED, 2019).

In essence, summarization can be viewed as a *selection problem*[1] or as a way to construct new data that represents the original source.

### 2.1.5 The Data Summarization Process

Data summarization, irrespective of data type, typically comprises four core components: (i) input data, the raw data to be summarized; (ii) preprocessing, an optional step that prepares the input data for summarization; (iii) summarization, the central task where various methods and approaches are employed to generate summaries; and (iv) summarized data, the result of the summarization process, presenting the reduced yet informative version of the input data. This generic process is observed in several works (BOUDIN; HUET; TORRES-MORENO, 2011; HEU; QASIM; LEE, 2015; MOHSIN et al., 2021; MA et al., 2022) and as depicted in Figure 2.

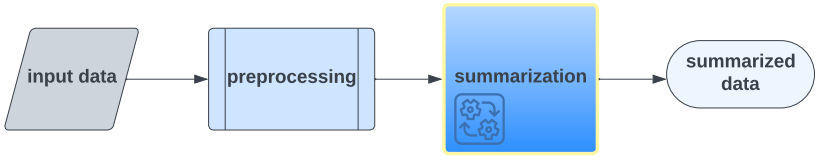


Figure 2 – Generic Process for Data Summarization

In conclusion, data summarization is a crucial tool for extracting valuable insights from large datasets. By aligning the choice of approaches and techniques with the specific goal of the summarization task, it is possible to empower decision-makers to navigate complexity effectively. This, in turn, enhances their understanding of intricate information domains, ultimately facilitating more informed and impactful decision-making processes.

[1] The selection problem consists of selecting the most appropriate elements of a predefined set of elements, i.e., the best ones from a given collection (DESU, 1970).

## 2.2 TRAJECTORY DATA

One of the foundational pillars of this work is the comprehensive exploration of trajectories of moving objects. With the widespread adoption of geolocation technologies and the ubiquity of tracking systems, trajectory data has become essential in various fields. In data analytics, trajectory data holds significant importance as it is increasingly being collected for mining, analysis, and decision-making (RENSO; SPACCAPIETRA; ZIMÁNYI, 2013).

Trajectory data, in essence, is a record of the movement of an object through the spatial and temporal dimensions. It is encountered in its simplest form as the *raw trajectory* - a sequential representation of the movement of an object across geographic space over time (ERWIG et al., 1999). This raw trajectory primarily consists of two fundamental dimensions:

- **spatial dimension**: This dimension encapsulates the geographic coordinates, such as latitude and longitude, precisely identifying the location of the object at distinct time intervals, i.e., its physical space;
- **temporal dimension**: This dimension refers to the timestamp, or a time intervals, associated with each spatial coordinate. This time information compose a chronological sequence, providing insights into the temporal aspects of the movement of an object and interactions.

Around 2007, the concept of *semantic trajectory* emerged, in which a third dimension is aggregated into data trajectories, i.e., a raw spatiotemporal trajectory *(x, y, t)* is enriched with semantic information. This third dimension is the semantic layer, which is infused with contextual information, such as a *point of interest (POI)* (*e.g.*, a restaurant) that the object had visited along its trajectory (ALVARES et al., 2007; PARENT et al., 2013).

This additional semantic dimension adds depth to trajectory data analysis and enables more meaningful insights. Consider Figure 3, which presents the trajectory of an individual during a single day. In this example, the raw trajectory maintains the spatiotemporal information about the individual (Figure 3(a)). Figure 3(b), in turn, shows a semantic trajectory with contextual information (POIs) associated with its points (home, work, and restaurant).

With the typical use of IoT and social media, enriching trajectories with a vast amount of semantic information has become possible. When trajectories or their individual points become associated with multiple and heterogeneous semantic contexts, they transform into what is known as *multiple aspect trajectories (MAT)* (MELLO et al., 2019). These MATs are characterized by the fusion of three dimensions: spatial, temporal, and semantic, where the semantic dimension may represent multiple and heterogeneous aspects.

Figure 3(c) shows the raw trajectory enriched with information like the mean of transportation used by the individual, postings on social networks, weather conditions, and so on. This example highlights that a multiple-aspect trajectory is a complex object whose attributes can hold simple or complex objects according to the context of each described domain.

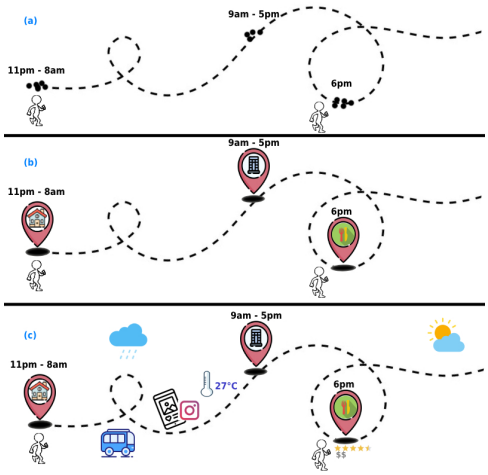


Figure 3 – An example of a raw trajectory (a), semantic trajectory (b), and a multiple aspect trajectory(c).

### 2.2.1 Applications and Challenges

Trajectory data finds its applications across various domains, including transportation and logistics (MARKOVIĆ et al., 2018; KONG et al., 2018), geographical phenomenons analysis (LEE; HAN; WHANG, 2007; ZHENG, 2015), location-based services (ZHENG, 2015; YANG; WANG; ZHANG, 2019; WANG et al., 2021), and social sciences (NARA, 2021). While trajectory data holds great promise, it also presents several challenges (MARTINEZ; CRISTOBAL; BELKOURA, 2018; GAO et al., 2019), such as:

- **Data Volume and Velocity**: Trajectory data can generate vast amounts of data, especially in scenarios involving numerous moving objects. Effectively managing this high data volume could be a complex task.

- **Complex Analysis**: Analyzing trajectory data demands advanced spatial, temporal, and semantic analytics, including trajectory clustering, anomaly detection, and predictive modeling. These analyses can be computationally intensive and require expertise in data science.

In the face of these challenges, data reduction is an alternative method to reduce the complexity of data management. The complexity of data management is mitigated by intelligently reducing the volume of trajectory data through techniques like trajectory summarization. This approach aims to combine similar trajectories and reduce the amount of data to be processed, making it more manageable for analysis while preserving essential patterns and insights.

## 2.3 TRAJECTORY SUMMARIZATION

*Trajectory data summarization* is a vital process that condenses extensive and complex trajectories into more manageable and informative summaries (ETIENNE et al., 2016). The main goals of trajectory summarization are:

- **Reducing Data Volume**: Managing large trajectory datasets can be challenging due to their vast volume, making storage and processing difficult (FENG; ZHU, 2016; GEORGIOU et al., 2018). Summarization techniques aim to reduce the volume of data while retaining essential information, making it more manageable.
- **Preserving Key Patterns**: Summarization techniques focus on capturing and preserving the key movement patterns and tendencies in the original data (AHMED, 2019). This ensures that valuable insights are not lost in the summarization process.

In essence, trajectory summarization involves a process that derives representative information from a dataset, often given by a representative selection/computation problem.

### 2.3.1 Representative Trajectory Data

The concept of a representative trajectory is pivotal in trajectory summarization. According to (LEE; HAN; WHANG, 2007; AYHAN; 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.

Approaches to determining representative data from a trajectory dataset can be broadly categorized into two types: those that compute a representative trajectory through mathematical computation (LEE; HAN; WHANG, 2007; ETIENNE et al., 2016; BORKOWSKI, 2017; GAO et al., 2019) and those that select specific trajectories or segments to represent the entire dataset (PANAGIOTAKIS; PELEKIS; KOPANAKIS, 2009; PANAGIOTAKIS et al., 2012), referred to as a *selection problem*.

Given a set of trajectories $\mathbf{T}$, the challenge lies in summarizing these data to obtain a compact yet informative representation, facilitating simplified analysis and decision-making processes. Thus, the problem of computing representative data involves deriving a trajectory that captures the essential characteristics of the original dataset while minimizing information loss.

For example, Figure 4(a) showcases individual MATs capturing various actions over several days (Sunday, Monday, and Tuesday). In contrast, Figure 4(b) illustrates a representative MAT computed through a summarization process applied to these individual MATs. The representative MAT effectively emphasizes frequently occurring actions (MACHADO; MELLO; BOGORNY, 2022a).

Analyzing trajectory patterns offers valuable insights for data analysts, enabling better decision-making. For instance, with representative trajectories, a recommendation system can learn the patterns of an individual and provide personalized recommendations. As demonstrated in Figure 4 (c), the system can identify a vegetarian restaurant along a new trajectory for the individual and recommend it to him/her.

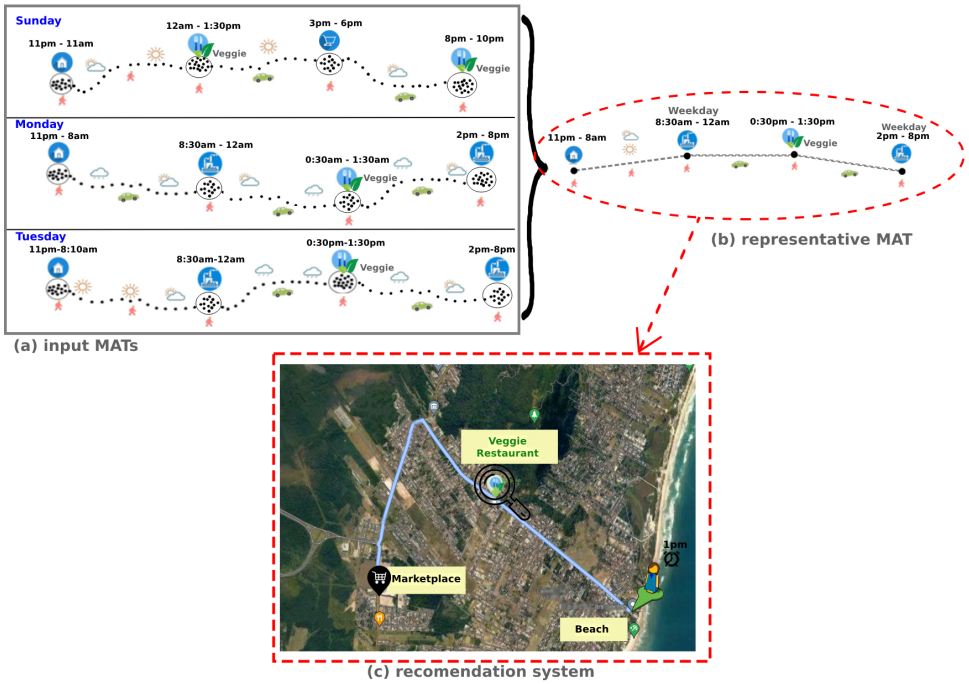


Figure 4 – An example of MATs (a), a representative MAT for them (b), and an example of recommendation based on its representative MAT (c). Adapted from Machado, Mello e Bogorny (2022a).

In summary, trajectory summarization is pivotal in handling and extracting insights from trajectory data, reducing their complexity while preserving essential information for various applications, i.e., the problem can be formalized as finding *RT*, as stated in *Problem 1*. To achieve this, resolving conflicts among similar data instances is essential, necessitating the application of *similarity measures* to identify data similarities. These measures are essential for analyzing trajectory data, including tasks such as clustering (LEE; HAN; WHANG, 2007), classification (PORTELA; CARVALHO; BOGORNY, 2022), and k-nearest neighbor search (SEEP; VAHRENHOLD, 2021).

### 2.3.2 Similarity Measures

Similarity measures are essential tools in trajectory data analysis, providing a basis for solving various analytical problems. These measures evaluate the similarity between trajectories and enable quantitative comparisons. There are several categories of similarity measures, depending on the focus of the analysis. In the following, we explore some of the most prominent ones.

- **Similarity measure based on spatial dimension**: The most popular category of similarity measures that rely exclusively on the concept of space (WANG et al., 2013), involving computing the topology and geometry, like *homotopy type* (BUCHIN et al., 2013) or computing distances between geographic coordinates, like *Euclidean distance*, *Haversine distance* and/or the *Hausdorff distance*. They assess how close trajectories are in physical space;

- **Similarity measure based on sequence data**: Designed for sequence data, these measures can be adapted for trajectory analysis. Well-known examples include *Dynamic Time Warping (DTW)* (BERNDT; CLIFFORD, 1994), *Longest Common Subsequence (LCSS)* (VLACHOS; KOLLIOS; GUNOPULOS, 2002), as well as the *Edit Distance (ED)* (SU et al., 2020). These measures focus on aligning trajectory points over time;

- **Similarity measure based on Temporal dimension**: Some measures are designed to consider the temporal dimension. They align timestamped locations in trajectory context, often associating spatial distance metrics with temporal similarity. Examples include *equal-time* and *similar-time* distance (BUCHIN; KILGUS; KöLZSCH, 2018);

- **Similarity Measure for Raw Trajectory**: Tailored for raw trajectories, these measures offer solutions that account for both spatial and temporal dimensions. Examples comprise the *Discrete Fréchet distance (DF)*, as a discrete variant of *Fréchet distance* (EITER; MANNILA, 1994), *SDist* (YING; XU; YIN, 2009), *Minimum Euclidean Horizontal (MEH)* distance (FRENTZOS et al., 2007) and *Uncertain Movement Similarity (UMS)* (FURTADO et al., 2018);

- **Similarity Measure for multidimensional in trajectory data**: These measures support all three trajectory dimensions: space, time, and semantics. Examples include *Multidimensional Similarity Measure (MSM)* (FURTADO et al., 2016), *Stops and Moves Similarity Measure (SMSM)* (LEHMANN; ALVARES; BOGORNY, 2019), and *Multiple aspect trajectory similarity (MUITAS)* (PETRY et al., 2019).

These similarity measures are crucial in summarizing trajectory data, as they allow for the analysis of similar trajectories and the computation of representative data. This is necessary to effectively summarize trajectory data. Such measures enable data analysts to gain insights into the movement patterns of objects and individuals across various domains.

# 3 RELATED WORK

To gain insight into the research issue, this chapter reviews existing surveys in the literature concerning trajectory data, specifically focusing on understanding the current challenges and open issues in this domain. These surveys shed light on problems related to storing and processing trajectory data, underscoring the need for effective solutions to mitigate these challenges. One such solution is reducing trajectory data by computing representative data using summarization methods. However, these problems become more complex with MATs. Therefore, we also review and analyze approaches related to trajectory data summarization, which involves reducing trajectory data to compute representative data.

## 3.1 SURVEYS ON TRAJECTORY DATA

In recent years, trajectory data has gained significant attention, showing multiple surveys exploring various topics. In this section, we analyze surveys published between 2016 and 2023 to identify trends and highlight the importance of understanding trajectory data. This analysis can guide future research. Table 1 depicts a comparison of these studies.

In our exploration, we have identified several surveys investigating trajectory data mining, data management, visual analytics, privacy, and data analytics. Specifically, seven surveys (ZHENG, 2015; FENG; ZHU, 2016; GEORGIOU et al., 2018; BIAN et al., 2019; da SILVA; PETRY; BOGORNY, 2019; XIE et al., 2020) focus on *trajectory data mining* using different methods, such as classification (BIAN et al., 2019; da SILVA; PETRY; BOGORNY, 2019), and prediction (GEORGIOU et al., 2018; XIE et al., 2020; HUANG et al., 2022; YIN; WEN; LI, 2023). In contrast, three surveys focus on *trajectory data management* (RICHLY, 2018; SU et al., 2020; WANG et al., 2021), but only one (SU et al., 2020) of them mentions query processing and similarity measurement. Additionally, one of them focuses on *visual analysis* (AHMED et al., 2019),

*Privacy* has emerged as a salient concern in trajectory data research, as noted in the survey by (FIORE et al., 2020). It underscores the need for risk assessments regarding attribute linkage and emphasizes anonymization as a pivotal process in trajectory data privacy protection.

Additionally, two surveys encompass *data analytics* (KONG et al., 2018; ALMEIDA et al., 2020), offering a more comprehensive view of research conduct and identifying key techniques and challenges in the field.

As we delve into the challenges and open issues identified across these surveys, *trajectory data management* surfaces as a recurring topic (ZHENG, 2015; FENG; ZHU, 2016; GEORGIOU et al., 2018; XIE et al., 2020; RICHLY, 2018; WANG et al., 2021; AMIGO et al., 2021; ALMEIDA et al., 2020). *Data volume*, particularly in the context of Big Data, stands out as a primary challenge (FENG; ZHU, 2016; GEORGIOU et al., 2018; AMIGO et al., 2021), encompassing storage, processing, and transmission. Scalable solutions for handling vast trajectory datasets remain a pivotal area of exploration (RICHLY, 2018; ALMEIDA et al., 2020).

Table 1 – Comparison of Surveys on Trajectory Data

| Search Area | Survey | Challenges and Open Issues | Contributions |
|---|---|---|---|
| Data Mining | (ZHENG, 2015) | - Big Data management<br>- Big data preparation<br>- Data representation<br>- Data mining | - Systematic review on data mining<br>- Analysis of methods to transform trajectories into other data formats |
| | (FENG; ZHU, 2016) | - Big Data management<br>- Understanding the behaviour of trajectories<br>- Privacy-preserving methods | Framework architecture for data mining |
| | (GEORGIOU et al., 2018) | Big data management and Prediction | - Formal definitions related to prediction<br>- Taxonomy of the solutions<br>- Properties of the datasets for validation purposes |
| | (BIAN et al., 2019) | Big data preparation and classification | Comparison of datasets using different classifiers |
| | (da SILVA; PETRY; BOGORNY, 2019) | - | Classification of clustering techniques |
| | (XIE et al., 2020) | Big data management and Prediction | Comparison of datasets using different prediction techniques |
| | (HUANG et al., 2022) | - Data limitation<br>- Prediction<br>- Ethical and Legal Considerations | - Analysis of popular prediction methods |
| | (YIN; WEN; LI, 2023) | - Data limitation<br>- Prediction<br>- Ethical and Legal Considerations | - Analysis and comparison of prediction methods |
| Data Management | (RICHLY, 2018) | Big Data management | Framework architecture for data mining |
| | (SU et al., 2020) | – | Classification and analysis of distance measures |
| | (AMIGO et al., 2021) | - Big Data management<br>- Understanding the behaviour of trajectories | Overview of reduction trajectory data, from compression to segmentation techniques |
| | (WANG et al., 2021) | Big data preparation and management | Data management overview |
| Visual Analytics | (AHMED et al., 2019) | Use of trajectory data in monitoring | Summary of trajectory data, public video data sets and methods for reducing footage |
| Privacy | (FIORE et al., 2020) | - Risk assessments of attribute linkage<br>- Anonymizing trajectory data | Research on privacy of trajectory micro-data |
| Data Analytics | (KONG et al., 2018) | - Privacy-preserving methods<br>- Understanding the behaviour of trajectories | Classification of trajectory data |
| | (ALMEIDA et al., 2020) | - Big Data management<br>- Understanding the behaviour of trajectories<br>- Privacy-preserving methods | Surveys on Big Data trajectory analytics with a focus on integration, design, and analysis |

Several surveys highlight the importance of *data reduction* and *data preparation* techniques (ZHENG, 2015; BIAN et al., 2019; GEORGIOU et al., 2018; FIORE et al., 2020; XIE et al., 2020; WANG et al., 2021; AMIGO et al., 2021). These methods aim to enhance data quality, reduce data volume, and automate tasks like data cleaning, ultimately facilitating more efficient analysis.

*Privacy protection* for trajectory data emerges as a critical concern, emphasizing the challenges related to anonymization (FIORE et al., 2020; KONG et al., 2018; ALMEIDA et al., 2020). *Understanding the behavior of trajectories* is another recurring theme, as it directly influences data analysis (ZHENG, 2015; FENG; ZHU, 2016; KONG et al., 2018; ALMEIDA et al., 2020; AMIGO et al., 2021). In addition, Huang et al. (2022), Yin, Wen e Li (2023) have identified data limitation as a significant challenge in the context of *trajectory prediction*. The availability, quality, and diversity of historical data are critical factors in obtaining high-quality training data. These studies emphasize the importance of obtaining representative trajectory

data to ensure accurate predictions.

Notably, trajectory data management and mining domains have received significant attention in recent surveys, with challenges related to data management. In this context, we highlight the challenges pointed out as the efforts to reduce the volume of data stored by data preprocessing tasks, aiming to improve data quality and minimize data size (RICHLY, 2018; WANG et al., 2021; FENG; ZHU, 2016; BIAN et al., 2019; AMIGO et al., 2021).

While reducing trajectory data is widely acknowledged as a challenge in data management, in-depth explorations are limited. Nevertheless, only two studies have explored this topic: Almeida et al. (2020), which focuses on data integration, unifying different sources into a single data format, and Amigo et al. (2021), which provides an overview of trajectory data compression. Amigo et al. (2021) consider several approaches to reducing a single trajectory into a more compact version, focusing on analyzing compression techniques and computing semantic knowledge. This thesis attempts to fill existing literature gaps, highlighting the data reduction challenges. In this way, an under-explored yet crucial theme refers to the *summarization of trajectory data*, so we also offer a comprehensive overview of the *state-of-the-art in trajectory data summarization*. The subsequent section presents a detailed exploration of relevant literature on this subject.

## 3.2 RELATED WORKS ON TRAJECTORY SUMMARIZATION

Trajectory data reduction is essential for refining complex trajectory datasets into manageable and informative representations. Research in trajectory data reduction, aimed at generating representative data, has seen significant advancements over the years. By analyzing and classifying related works, we have identified eleven studies focused on data summarization.

To provide a comprehensive understanding of the landscape, we begin with an overview of the related works, followed by an analysis categorized into different topics. These topics encompass *(i) representative data type, (ii) methods performed by the approach*, and *(iii) evaluation of the approach*.

Our research into trajectory summarization begins with related works dating back to 2007. A pioneering contribution by **Lee, Han e Whang (2007)** addressed the challenge of processing trajectory data by proposing a partition-and-group framework for spatial trajectories. Their approach involved two steps: partitioning trajectories using an approximation algorithm and then clustering the segments using *TRAjectory CLUStering (TRACLUS)* density-based clustering algorithm, which is based on the clustering algorithm *Density-Based Spatial Clustering of Applications with Noise (DBSCAN)* (ESTER et al., 1996). The proposed TRACLUS algorithm demonstrated its effectiveness in summarizing trajectory data by identifying common behaviors among subtrajectories, i.e., a common subtrajectory is defined as the representative trajectory for each cluster, which can be identified as the summarized data for this set of subtrajectories.

In 2012, **Panagiotakis et al. (2012)** contributed to this progress with their work. Their study focused on spatiotemporal trajectory segmentation and sampling in Moving Object

Databases (MOD) to capture shared portions between trajectories. They segmented trajectories into subtrajectories, clustered them to identify shared ones, and selected the most representative subtrajectory for each cluster based on density and similarity. The study proposed a method for capturing shared subtrajectories and selecting representative subtrajectories, improving trajectory data summarization. Their approach effectively represented trajectories by capturing shared portions and selecting representative subtrajectories by both spatial and temporal dimensions.

In 2013, **Buchin et al. (2013)** made a noteworthy contribution to trajectory data summarization. Their study aimed to compute the median trajectory in a set of input spatial trajectories, considering both *simple median* and *homotopic median* approaches. They segmented input trajectories into subtrajectories, arranged them, and determined the middle trajectory as a representative using either simple or homotopic median computation. By considering obstacles in the route, the study improved the understanding of trajectory paths, and introducing both the *simple median* and *homotopic median* methods offered flexibility in trajectory summarization. The comparison of the two approaches revealed that the *homotopic median* outperformed the *simple median* in most cases, highlighting its effectiveness in trajectory summarization and demonstrating advancements in this field.

In 2015, a significant advancement was made by **Ayhan e Samet (2015)**, who introduced DICLERGE (Divide-Cluster-Merge), a novel clustering framework for spatiotemporal trajectories designed explicitly for aircraft trajectory data. DICLERGE divides trajectories into three major flight phases (climb, enroute, and descent) and clusters each phase separately. The framework also includes the generation of a representative trajectory, achieved through lateral and vertical smoothing processes. Lateral smoothing involves filtering and connecting the cluster centroids of the trajectory points, while vertical smoothing determines the enroute altitude by calculating the median altitude of all trajectory points. DICLERGE offers a tailored approach to divide, cluster, and summarize aircraft trajectories into distinct flight phases, providing representative data for this specific context.

In 2016, **Etienne et al. (2016)** made a notable contribution by introducing a novel method for describing the typical movement of a cluster of homogeneous spatiotemporal trajectories. The study addressed the challenge of summarizing the central tendency for such clusters by the introduced method of Trajectory Box Plot (TBP). The TBP computes a representative trajectory by selecting an initial reference trajectory and pairing positions within it with corresponding positions in other trajectories within the same cluster. The central position for each cluster point is computed, and these ordered central positions are connected to generate a new reference trajectory. This iterative process repeats until the reference trajectory converges to a central (representative) trajectory. The computation process is based on the work of Petitjean, Ketterlin e Gançarski (2011). The study introduced the Trajectory Box Plot (TBP) to represent the typical movement of homogeneous spatiotemporal trajectories by central tendencies.

In 2018, **Agarwal et al. (2018)** addressed the challenge of subtrajectory clustering within spatial trajectories. The goal was to cluster subsequences of trajectories effectively to capture shared portions, identify segments with shared characteristics among trajectories, and

provide a summarized representation of the trajectories. The ultimate goal was to find the optimal set of subtrajectories that could effectively represent the entire input dataset. The study utilized an approximation algorithm based on the *Set-Cover problem* (CORMEN et al., 2009), an effective approach to this challenge to compute the representative trajectory for each cluster.

In 2019, **Gao et al. (2019)** presented a compression model for spatiotemporal trajectories enriched with semantic information. Their study aimed to improve the representation and compression of trajectory data by incorporating semantic aspects. The authors introduce a multi-resolution synchronization-based clustering model called *CascadeSync*. This model identifies delimited regions of geographic space, referred to as *Region of Interest (ROI)*, by clustering raw trajectory points. Gradually, these clusters are synchronized hierarchically, leading to the formation of a hierarchical ROI network. This process reduces the number of ROIs as the area size of each region increases. The study introduced a novel approach where each original trajectory can be compressed into a sequence of ROIs, incorporating semantic information. The approach of using CascadeSync for hierarchical ROI clustering and incorporating semantic information into trajectory compression demonstrated advancements in handling spatiotemporal trajectories enriched with semantics.

In 2020, the authors complemented their 2019 study by proposing a hierarchical embedding model. This model allowed the incorporation of each ROI/trajectory as a continuous vector in a semantic vector space. Significantly, it facilitated semantic similarity computation between two ROIs/trajectories through Euclidean distance metrics (GAO et al., 2020).

Additionally, in 2019, **Buchin, Kilgus e Kölzsch (2019)** introduced a framework called *Group Diagram (GD)* for representing spatiotemporal trajectories. The framework aims to represent input trajectories with minimal subtrajectories while preserving their essential characteristics. It generates a single representative trajectory called the *minimal GD*. The minimal GD is computed through a segmentation step, where subtrajectories within the input trajectories are clustered. The representative subtrajectory was computed as the middle subtrajectory for each cluster, considering a predefined maximum distance from all other subtrajectories. These representative subtrajectories are then connected to form the representative trajectory. The GD framework provides a novel data representation for spatiotemporal trajectories and offers an approach to summarize trajectories while maintaining essential information. Using the GD framework and the approximation algorithm based on the Set-Cover problem facilitated the generation of representative trajectories, demonstrating advancements in trajectory summarization techniques.

**Seep e Vahrenhold (2019)** proposed a solution for generating representative semantic trajectories also in 2019. Their study focused on identifying a sequence of transitions common to most routes within trajectory data, aiming to capture essential trajectory patterns. The authors considered a Finite State Machine (FSM) version called *Extended FSM (EFSM)* for their approach. In EFSM, each state represents a data point, and the sequence of states and transitions constitutes a subtrajectory. The complete sequence of states and transitions generated the representative trajectory, which captured common patterns among trajectories. The common pattern

was defined by analyzing the routes along the time dimension, and each transition was defined depending on the spatial and non-spatial aspects (SEEP; VAHRENHOLD, 2021). While specific findings and results were not mentioned in the short paper format, using EFSM to infer representative semantic trajectories demonstrates advancements in trajectory analysis.

In 2021, the authors advanced their work by proposing a method called *EFSMClust*, which extends the k-means algorithm to cluster trajectories with multiple aspects (SEEP; VAHRENHOLD, 2021). The clustering algorithm defines a similarity measure between a trajectory and a graph-based representation of a cluster centroid. The computed representative trajectory based on EFSM in the previous work is used to define the centroid of each cluster, and by using the similarity measure, they define the trajectories nearest to each representative trajectory to compute the final cluster. As this second work refers to another part and does not focus on representative computation, this step is not detailed. Although the study does not provide specific findings or results in the short paper format, it demonstrated advancements in trajectory analysis by using representative trajectories to improve cluster computation.

In 2020, **Rodriguez e Ortiz (2020)** introduced an approach for generating a representative trajectory from spatiotemporal trajectories. The study aimed to identify and represent patterns within trajectory data, effectively summarizing the underlying information. The input trajectories were initially segmented into subtrajectories, breaking them into more manageable parts. Subtrajectories were grouped, and pattern detection was performed using the DBSCAN algorithm, similar to the approach in the study by Lee, Han e Whang (2007). A representative trajectory is identified for each cluster using an arrangement of the spatial data of these subtrajectories. The study contributed to the field by introducing an approach that relied on pattern detection and spatial data arrangement to represent trajectories effectively.

**Li (2021)** proposed, in 2021, a method for extracting typical ship trajectories using Automatic Identification System (AIS) data and trajectory clustering. The study aimed to identify and provide representative ship motion trajectories from a set of ship trajectories. The study involved preprocessing AIS data and preparing it for trajectory analysis. Ship trajectories were segmented into meaningful subtrajectories. The improved DBSCAN clustering algorithm was applied to cluster trajectories to identify their typical mobility. The result is representative trajectories using the center of the clusters. The study contributed to the field by introducing a method for extracting and representing distinct ship trajectories using an improved DBSCAN clustering algorithm.

Finally, in recent research, **Pugliese et al. (2023)** presented a novel approach for MAT summarization by enriching raw trajectories with semantic context. However, their approach does not consider input data trajectories with multiple aspects, as the raw trajectories are enriched during the process, and group representative data is created for each group rather than just one representative information.

### 3.2.1 Discussion

Based on the discussion in Section 2.2, the evolution of trajectory summarization for trajectory data shows that semantic trajectories emerged in 2007. However, there is still a need for further improvement in summarizing this data, especially within the context of MATs that involve multiple and complex aspects. The challenge is to improve the summarization of the semantic dimension, which only began in 2019. We present a comprehensive table (Table 2) to compare related works regarding representative data. This table examines critical elements of each study: *(i)* considered dimension; *(ii)* summarization type; and *(iii)* mapping information.

Table 2 – Related work comparison

| Study | Considered Dimensions[2] | | | Summarization type | Mapping information |
|---|---|---|---|---|---|
| | Space | Time | Semantic | | |
| **TraClus (LEE; HAN; WHANG, 2007)** | X | | | Computation | |
| **(PANAGIOTAKIS et al., 2012)** | X | X | | Selection | |
| **Median Trajectory (BUCHIN et al., 2013)** | X | | | Computation | |
| **DICLERGE (AYHAN; SAMET, 2015)** | X | X | | Computation | |
| **TBP (ETIENNE et al., 2016)** | X | X | | Computation | |
| **(AGARWAL et al., 2018)** | X | * | | Computation | |
| **CascadeSync (GAO et al., 2019)** | X | X | | Computation | X |
| **GD (BUCHIN; KILGUS; KöLZSCH, 2019)** | X | X | | Selection | |
| **(SEEP; VAHRENHOLD, 2019)** | X | * | * | Computation | |
| **(RODRIGUEZ; ORTIZ, 2020)** | X | * | | Computation | |
| **(LI, 2021)** | X | * | * | Selection | |
| **MAT-SG (ours)** | X | X | X | Computation | X |
| **MAT-SGT (ours)** | X | X | X | Computation | X |

[2] In the Considered Dimension column, "X" indicates a completely resolved dimension, "*" indicates a dimension that is not completely resolved, and an empty cell indicates a dimension that is not addressed in the study or is not mentioned.

First, some studies, as marked with (*), do not encompass all dimensions provided in the input trajectories within their methods. Observing the works that consider the semantic dimension, Seep e Vahrenhold (2019) consider a trajectory annotated with additional information (a semantic trajectory), where all attributes of the points are treated as a spatial or non-spatial value. Li (2021) refers to the vessel scenario and consider specific aspects of the semantic dimension (vessel speed and direction) while reducing trajectory points and computing representative data. Other works use semantic dimension only to enrich their final raw data, without regarding it as input data or incorporating it into their summarization method, as observed in Gao et al. (2019).

As expected, the spatial dimension is a consensus among all studies, and most of them also include the temporal dimension. However, only some studies deal with semantic aspects in

their processes to compute representative data, making this problem an open issue.

A notable observation is that while trajectory summarization can be categorized into two types: *computation* of representative data through mathematical methods or *selection* of specific trajectories or segments to represent the entire dataset, as described in Section 2.3.1, most related works primarily focus on the computation task. Over the years, this emphasis on computation has led to significant advancements in summarization techniques. However, it is essential to recognize that the *semantic dimension* has been somewhat overlooked in these computations, even in the most recent studies. While spatial and temporal dimensions are consistently addressed, capturing and representing all semantic aspects remains an open issue.

Furthermore, mapping information, which pertains to understanding the relationship between input trajectories and the summarized data, is another critical aspect that has received limited attention. Notably, only one study (GAO et al., 2019) describes this topic explicitly. Other related works do not perform mappings or do not provide any information about them.

In Gao et al. (2019), they propose a method to convert an input trajectory into an ROI network. Each ROI represents the origin of an input trajectory, and the trajectory itself can be represented as a sequence of ROIs within the ROI network. Furthermore, all trajectories passing through a particular ROI are recorded. However, this mapping information primarily indicates which input trajectory contributes to a particular ROI without specifying the specific points within the trajectory that form the ROI.

### 3.2.2 Evaluation of the Representative Trajectory Data Computation Process

This section conducts a thorough analysis of evaluations presented in related works, focusing on considered datasets and the evaluated factors. Special attention is given to the reduction process used for computing representative data, specifically in relation to the summarization task, as indicated by bold highlighting in the *Evaluated factor* column of Table 3. Additionally, for studies that explicitly evaluate the computation of representative data, we provide an overview of their evaluation.

Only three works evaluate their approach regarding some factor of computation of representative data (PANAGIOTAKIS et al., 2012; BUCHIN et al., 2013; SEEP; VAHRENHOLD, 2019), referring reduction process, as highlighted in bold in Table 3, and the evaluation of these studies are detailed in the following.

Among the related works, only five ((PANAGIOTAKIS et al., 2012; BUCHIN et al., 2013; SEEP; VAHRENHOLD, 2019), MAT-SG, and MAT-SGT) systematically evaluate their approach concerning the computation of representative data, focusing on the reduction process, as highlighted in bold in Table 3. The evaluations of these studies are detailed below.

In Panagiotakis et al. (2012), the approach undergoes a quantitative evaluation using both real and synthetic databases. The process involves computing representative data as sub-trajectories concentrated at the center of the cluster. The evaluation focuses on determining optimal parameters for clustering trajectories. To construct the sub-trajectory sampling set, a

Table 3 – Comparative of the related works w.r.t. evaluated factor

| Study | Dataset | Evaluated factor | Compares to |
|---|---|---|---|
| TraClus (LEE; HAN; WHANG, 2007) | Hurricanes[3] and Animals[4] (Elk and Deer) | Clustering process | - |
| (PANAGIOTAKIS et al., 2012) | Synthetic dataset, Transport[5] (Athens trucks), Transport[6] (Milano) | **Performance sampling Representative data** | standard sampling (random and stratified sampling) |
| Median Trajectory (BUCHIN et al., 2013) | Synthetic dataset | **Computed medians** | - |
| DICLERGE (AYHAN; SAMET, 2015) | Private dataset (Aircraft) | Clustering process | - |
| TBP (ETIENNE et al., 2016) | Vessels[7] (AIS Brest, France) | - | - |
| (AGARWAL et al., 2018) | Synthetic data, Geolife[8] Urban Taxi (Beijing) (LIAN; ZHANG, 2018), Private dataset | Clustering process | - |
| CascadeSync (GAO et al., 2019) | Synthetic dataset, Geolife, Hurricanes, Urban Taxi[9] (T-Drive) and Animals[10] (Barn Swallows) | Compression algorithm | - |
| GD (BUCHIN; KILGUS; KöLZSCH, 2019) | Animals[11] (LifeTrack Geese) | Data representation process | - |
| (SEEP; VAHRENHOLD, 2019) | Hurricanes and Geolife | **Representative data** | TRACLUS and Median Trajectory |
| (RODRIGUEZ; ORTIZ, 2020) | Private dataset (Transport) | - | - |
| (LI, 2021) | Private dataset (Vessel) | Clustering processing | - |
| MAT-SG (ours) | Private dataset (Pisa), Foursquare[12], Brightkite[13], Gowalla[14] | **Representative data** | MAT-SGT |
| MAT-SGT (ours) | Private dataset (Pisa), Foursquare, Brightkite, Gowalla | **Representative data** | MAT-SG |

* Highlighted in bold are the evaluated factors applied to the computation of the representative data.
[3] http://www.nhc.noaa.gov/data/hurdat/
[4] https://www.fs.usda.gov/research/pnw/forestsandranges/locations/starkey
[5] http://www.chorochronos.org/Default.aspx?tabid=71&iditem=31
[6] Milano dataset consists of GPS traces describing the movement of a set of 17K vehicles during one week at the beginning of April 2007 (not available).
[7] http://www.chorochronos.org
[8] https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/
[9] https://www.microsoft.com/en-us/research/publication/t-drive-trajectory-data-sample/
[10] https://www.datarepository.movebank.org/handle/10255/move.655
[11] https://zenodo.org/records/3508780
[12] https://github.com/bigdata-ufsc/datasets_v1_0/tree/main/data/multiple_trajectories/Foursquare_NYC
[13] https://github.com/bigdata-ufsc/datasets_v1_0/tree/main/data/multiple_trajectories/Brightkite
[14] https://github.com/bigdata-ufsc/datasets_v1_0/tree/main/data/multiple_trajectories/Gowalla

proposed distance measure calculates the number of trajectories in the input dataset represented in each sampling. For comparative analysis with other sampling techniques, the *Root Mean Square Error (RMSE)* metric is employed, indicating which technique offers superior coverage of the space-time within the input dataset.

In Buchin et al. (2013), two approaches, namely the simple median and homotopic median, are systematically compared through both quantitative and qualitative analyses. The quantitative evaluation involves considering metrics such as the number of vertices, total length, and total turning angle of the median trajectory computed by both approaches, along with the average of these measures for the input trajectories. Qualitative analysis is conducted through

visual inspection, revealing that the homotopic median results better results.

In the work by Seep e Vahrenhold (2019), a comprehensive quantitative and qualitative evaluation is conducted on its representative data, comparing results against TRACLUS and Median Trajectory. The qualitative assessment involves visual analysis, while the quantitative evaluation employs means and median distance (using Fréchet distance) between input and representative data for each approach. This work demonstrates a more faithful representation in both evaluations of the considered datasets.

### 3.2.3 Summary

Since summarizing trajectories is a vital process that condenses extensive and complex trajectories into more manageable and informative summaries, and while MATs have emerged as a promising data type, offering extensive possibilities for data analysis, it is noteworthy that state-of-the-art approaches that summarize MATs tend to overlook the consideration of all semantic aspects individually. Additionally, there is a lack of studies that summarize MATs and provide mapping information regarding the relationship between the representative data and corresponding input points. This information would allow us to understand the origin of each part of our representative data.

In response to these observations, this thesis introduces two novel approaches, MAT-SG and MAT-SGT, designed to address these challenges. In the following chapter, we delve into the details of these approaches, elucidating their methodology and contributions to the trajectory summarization field.

Furthermore, an examination of related works reveals that only one study (SEEP; VAHRENHOLD, 2019) encompasses all three dimensions (spatial, temporal, and semantic) in evaluating the representative trajectory computation process. However, it is crucial to note that this evaluation primarily relies on visual analysis (qualitative evaluation), with quantitative aspects limited to the spatial dimension. The authors mention the lack of a well-defined measure for quantitative evaluation and assessing the degree to which the representative data genuinely represents all the input data.

In response to this gap, we propose a novel representativeness measure (RMMAT) detailed in Chapter 5. This measure aims to provide a robust quantitative evaluation measure, addressing the identified need for a comprehensive assessment of how well representative data truly represents all aspects of the input trajectory data. Only the study by Seep e Vahrenhold (2019) compares their work with methods that encompass spatial dimensions (LEE; HAN; WHANG, 2007; BUCHIN et al., 2013), but no quantitative comparison is performed. The datasets used in Seep e Vahrenhold (2019) refer to spatial (Hurricanes) and semantic (Geolife) analysis. However, the analysis of the Geolife dataset is only qualitative, focusing on patterns identified by specific user. Since no compatibility baseline is available, we choose to use datasets involving MATs, providing quantitative evaluation as detailed in Chapter 6.

# 4 METHODS FOR MULTIPLE ASPECT TRAJECTORY DATA SUMMARIZATION

In the fast-paced world of data management, the challenge of reducing trajectory data to improve data processing and data mining is essential. As we explored in Chapter 2, this challenge remains a little explored, and in Chapter 3, we examined related works that emphasized summarizing trajectories, observing that primarily their focus is based on spatial and temporal dimensions. However, the semantic dimension, which holds the key to opening a deeper understanding of trajectory data, remains largely unexplored.

Considering that the semantic aspects provide context and meaning for both the object and its movement regarding its raw trajectory, MATs, with their multidimensional nature, possess the power to provide comprehensive insights into object movement and its associated aspects. However, these aspects usually are not considered in representative data, leading to a combinatorial explosion that requires additional summarization strategies, particularly for MATs computing representative data. These strategies could provide insights into both object movement and associated aspects.

In this chapter, we delve into strategies designed to confront this challenge head-on. These strategies offer summarization for computing representative data from MATs, harnessing the full spectrum of information contained within these trajectories. We present two novel methods developed during the Ph.D. research, each employing distinct strategies for reducing trajectory data through summarization.

The first method, detailed in Section 4.1, is named MAT-SG (MACHADO; MELLO; BOGORNY, 2022a), which computes representative data exploring the pattern involved in each spatial area regarding input MATs, effectively summarizing all involved aspects. MAT-SG stands as the pioneering trajectory summarization method explicitly tailored for MATs, addressing various dimensions while treating each aspect individually. However, it is important to note that MAT-SG may not provide optimal solutions when preserving temporal sequences within the representative data is crucial. For this reason, Section 4.2 introduces the second method, MAT-SGT (MACHADO et al., 2023a), which aims to provide representative data that capture the temporal sequences within input MATs, summarizing all related aspects.

These methods are presented to bridge the gap between MATs and the need for more efficient and informative summarization, providing valuable tools for data analysts and researchers in various fields.

## 4.1 MAT-SG: MULTIPLE ASPECT TRAJECTORY SUMMARIZATION BASED ON A SPATIAL GRID

The method outlined in this section is a significant contribution to this Thesis. It introduces a novel algorithm known as MAT-SG (*Multiple Aspect Trajectory Summarization based on a spatial Grid*), which is designed to reduce the input dataset while aspiring to offer representative data that encapsulates the predominant patterns within MATs. MAT-SG extracts essential

insights about the moving object, such as its spatial regions and relevant attributes associated with these regions. Our method focuses on computing a representative MAT that accurately reflects the primary behavior and characteristics of the input MATs, taking into account the spatial density and frequency of each attribute value. MAT-SG was developed to address the lack of summarization methods for reducing input MATs while still resulting in representative data.

We assume the input MATs were already filtered by some criterion[15]. So, the representative MAT denotes the primary behavior of these input MATs considering spatial density and frequency of each aspect attribute value.

### 4.1.1 Data model

To maintain representative MAT generated by MAT-SG, we rely on a conceptual data model shown in Figure 5. This conceptual model provides a standardized representation of the input data and keeps the representative points and their mappings to the input points. Each point, in turn, holds information about all dimensions: *space* (x and y coordinates), *temporal aspects* (that could be represented by a timestamp or a time interval, denoting the start and end times), and *semantic aspects* (a set of the attributes with their corresponding values). Each attribute belongs to a categorical or numerical data type.

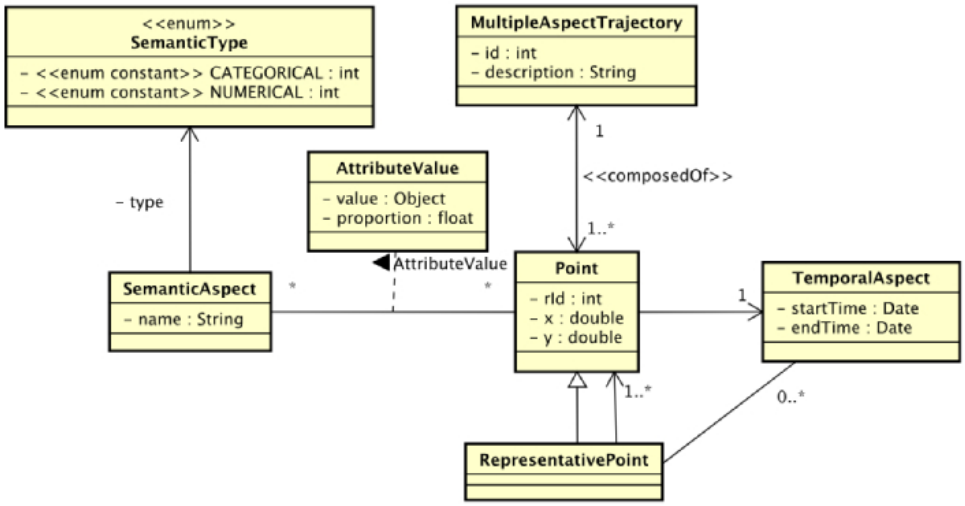


Figure 5 – The conceptual model for MAT-SG.

This Thesis introduces a concept that contributes significantly to the model for representative data. This model enables mapping data between input MATs and the resultant representative MAT. The representative MAT is structured as a set of *representative MAT points*, denoted as $p_r$. The MAT-SG algorithm computes these points, and the representative MAT ($RT$) is essentially composed of a sequence of $p_r$'s.

[15] These criteria are out of the scope of this paper, but examples could encompass operations like clustering or straightforward filtering. For example, these criteria might involve tasks such as given MATs generated by check-ins of different individuals to discern their patterns during specific time periods. A simple filter could ensure that the dataset contains only the trajectories of a particular individual during these defined time intervals.

To compute $RT$, we summarize the information into $p_r$'s. Each $p_r$ summarizes relevant information derived from multiple input MAT points, and a relationship is established and maintained between $p_r$ and its corresponding MAT points to ensure accurate representation. It is important to note that $p_r$ is a specialized MAT point that preserves specific attributes, contributing to its significance in the summarization process. This attribute-holding capacity of $p_r$ further enhances its value in representing and maintaining crucial information within MATs. Unlike input MAT points, each $p_r$ provides a set of Temporal Aspects, as the representative point represents a representative spatial region and the usual activities in that region, including the usual time that the object frequents each region.

### 4.1.2 Architecture

An overview of the MAT-SG method is presented in Figure 6, illustrating its core component: *Data Summarization*, which comprises two main sub-components: *(i) Data Segmentation* and *(ii) $p_r$ computation*. The first one, Data Segmentation, aims to discern underlying data patterns based on spatial density. The second one, $p_r$ Computation, is focused on summarizing data by analyzing its frequency.

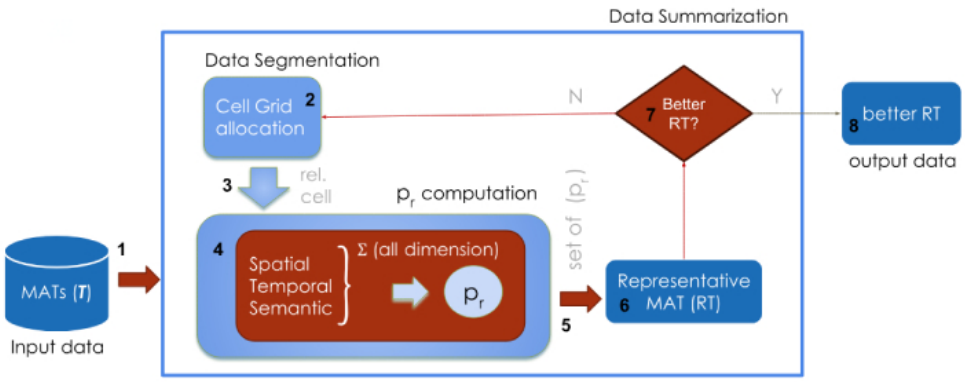


Figure 6 – MAT-SG overview.

The process begins with an input set of filtered MATs (**T**) in step 1. These MATs are selected based on specific criteria, although the specific details of these selection criteria are not discussed in this thesis. We assume that **T** exhibit some degree of similarity among the selected MATs. Step 2 involves segmenting the input MAT points into a spatial cell grid, facilitating the identification of relevant cells. For each of these relevant cells, step 4 is performed to calculate representative points that comprehensively summarize all dimensions and encapsulate the essential characteristics of the input data within each cell.

The outcome is the group of all computed representative points within a MAT object, resulting in the $RT$as the output data in steps 5 and 6. To refine the results, step 7 involves the selection of the best among the computed $RT$*s* as the final output. MAT-SG provides a comprehensive representation of the primary behaviors and characteristics demonstrated by the

input MATs, taking into account spatial density and the frequency of each aspect attribute value. The following section will provide a detailed exploration of the MAT-SG process.

### 4.1.3 Algorithm

MAT-SG considers a set of input parameters besides the input MATs. They are detailed in Table 4. $\tau_{rc}$ and $\tau_{rv}$ are optionally defined by the analyst; otherwise, default values are assumed. MAT-SG starts by calculating $rc = |T.points| \times \tau_{rc}$, which is based on a proportion $\tau_{rc}$. For example, given $\tau_{rc} = 1\%$ and $|T.points| = 200$, then $rc = 2$. In other words, only cells with a minimum of 2 points are considered relevant for accommodating a $p_r$. Subsequently, MAT-SG proceeds through its steps, meticulously detailed in the following sections.

Table 4 – Parameters of our summarization methods

| Parameter | Explanation | Default |
|---|---|---|
| **T** | Set of previously filtered input MATs | - |
| $\tau_{rc}$ | Minimum proportion of all input MAT points $\|T.points\|$, deciding if a cell is considered a relevant cell to compute $p_r$ | $rc = 2$ |
| $\tau_{rv}$ | 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-SG algorithm, detailed in Algorithm 1, is designed to compute an optimal *RT* by identifying the most suitable spatial segmentation. It initiates by determining the *minimum spatial threshold* ($\tau_s$) to measure the dispersion among all input points. Subsequently, it calculates the distance between the grid origin (0,0) and the farthest point from it (line 6). This calculation helps establish the maximum grid size, assuming all points fall within a single cell. The initial $z$ value computed in this process is a multiplier for determining the cell size. Using this initial $z$ value, the algorithm creates an initial grid with a single cell encompassing all MAT points, configuring the start of the process (line 11 - more detailed in Algorithm 2). The subsequent steps involve iteratively reducing the $z$ value to analyze and compute an improved *RT* (lines 10 to 25). This iterative approach aims to identify the optimal segmentation that yields the most refined *RT*. Additionally, the option to reduce $z$ value in 15% in each interaction (line 25) was determined after conducting various tests to identify the most effective reduction rate, considering both runtime efficiency and the sufficiency of information for generating a new *RT*.

Within each iteration, the algorithm performs data spatial segmentation based on the current $z$ value, culminating in spatial allocation (*Cell Grid allocation* step). Subsequently, it computes representative points for each group of points (lines 12 to 14). As previously mentioned, MAT-SG accomplishes MAT summarization through two key internal components: *(i) data segmentation*; and *(ii)* $p_r$ *computation*.

To assess the quality of the computed *RT*, it is compared to the previously calculated representative trajectory (*betterRT*), with a stipulated margin of 10% improvement. If a superior

*RT* is identified, the algorithm updates the *betterRT* and resets the counter-tracking iterations without improvement. The best *RT* is determined by its similarity, coverage, and superiority over others in two new computations. Section 4.1.3.3 provides a detailed explanation of the selection process.

**Algorithm 1:** MAT-SG

```
   input  : T, τ_rc, τ_rv
   output: RT                                   /* representative trajectory */
 1 rc ← |T.points| × τ_rc;
 2 τ_s ← computeMinSpatialThreshold();
 3 rt ← ∅
 4 betterRT ← ∅
 5 count ← 0;
 6 z ← computeMaxZValue();
 7 betterRTmeasure ← 0;
 8 w_sim ← 0.5;
 9 w_cover ← (1 − w_sim);
10 while z > 1 do
       // component (i) - Fig. 6 (steps 2 and 3)
11     spatialCellGrid ← cellGridAllocation(rc, z, T) // Algorithm 2
       // component (ii) - Fig. 6 (step 4)
12     foreach eachGroupPoint ∈ spatialCellGrid do
13         p_r ← computeRepPoint(eachGroupPoint, τ_rv);
14         rt ← rt ∪ p_r // Fig. 6 (step 5)
       // analysis of better RT - Fig. 6 (step 7)
15     rtMeasure ← RMMAT(rt, T, w_sim, w_cover);
16     if (rtMeasure × 1.1) ≥ betterRTmeasure then
17         betterRTmeasure ← rtMeasure;
18         betterRT ← rt;
19         rt ← ∅
20         count ← 0;
21     else
22         count + +;
23     if count > 1 then
24         break;
25     z ← z × 0.85;
26 return betterRT;
```

The two components of the MAT-SG method are detailed next.

#### *4.1.3.1 Data Segmentation Component*

The initial step of the MAT-SG algorithm involves segmenting the points from the input MATs into a grid of square cells. This process is illustrated in Figure 7, which depicts a spatial grid with a highlighted cell. The size of each cell is determined by a threshold of spatial

dispersion ($\tau_s$), which specifies the maximum spatial distance between any two points within the cell. In other words, this threshold represents the diagonal length of each cell.

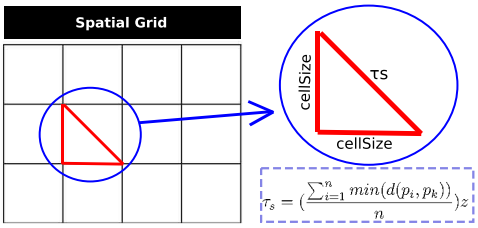


Figure 7 – Cell size computation

The calculation of $\tau_s$ (as shown in the equation in Figure 7) is performed dynamically and automatically. It is computed based on the average minimum spatial distances between the input MAT points. In the case of a given input set $\mathbf{T}$ with $n$ points, we calculate the Euclidean distance $d()$ for each point $p_i \in \mathbf{T}$ concerning its nearest neighbor $p_k \in \mathbf{T}$. The value of $\tau_s$ is then obtained by multiplying a factor $z$ with the average of these minimal distances.

The size of these grid cells (cell size) essentially determines the granularity of the spatial segmentation. Once the cell size is established, the input MAT points are allocated to the appropriate cells within the spatial grid. After this allocation, the process identifies the so-called *relevant cells*, which contain a sufficient number of points (at least $rc$) to provide meaningful representation and insights.

Algorithm 2 details the Data Segmentation step. An advantage of this approach is that it only generates cells that contain points, optimizing memory usage. It allocates the points from $\mathbf{T}$ to a spatial grid, implemented as an inverted index (or *inverted list*). In this implementation, the *key* represents the identity of the cell position, while the *value* comprises a list of the $\mathbf{T}$ points allocated within that cell.

**Algorithm 2:** MAT-SG:*cellGridAllocation*

```
   input  : rc, z, T
   output: spatialCellGrid                                   /* inverted list */
 1 spatialCellGrid ← ∅;
 2 τs ← computeτs(z);
 3 cellSize ← computeCellSize(τs);
 4 foreach T ∈ T do
 5     foreach p ∈ T do
 6         key ← getCellPosition(px, py, cellSize);
 7         if spatialCellGrid.get(key) = ∅ then
 8             spatialCellGrid.new(key);
 9             spatialCellGrid.get(key).put(p);
10         else
11             spatialCellGrid.get(key).append(p);
12 spatialCellGrid.updateByRelevantCells(rc);
13 return spatialCellGrid
```

The algorithm first computes the cell size of the spatial grid (lines 2 and 3) considering that all cells in the grid have the same size. To determine the appropriate cell for each MAT point, it calculates the grid position key for point $p$ (line 6), considering the cell size defined by

the *getCellPosition* function, which is calculated as $(\frac{p_x}{cellSize}, \frac{p_y}{cellSize})$. Then, $p$ is allocated into the grid cell of this position (lines 6 to 11), provided that the cell already exists.

In cases where the designated cell does not exist, the algorithm creates a new key and inserts the point into it (lines 8 and 9). This process of allocating points to grid cells ensures efficient spatial segmentation and allocation of the input MAT points while optimizing memory usage. The *spatialCellGrid* is updated to maintain only relevant cells (line 12). It identifies cells containing sufficient points (at least *rc*) to offer meaningful representation and insights.

The spatial data segmentation is efficiently done with the allocation method, which optimizes memory utilization by dynamically generating cells when necessary. Identifying relevant cells ensures that the subsequent algorithm step focuses on regions of interest with sufficient data density to provide valuable summarization and insights.

#### *4.1.3.2 Representative Point Computation component*

In the second component of MAT-SG, the aim is to summarize each group of points (*eachGroupPoint*) obtained from the first component. This is depicted in Figure 6 (step 4). It is important to note that cells containing less than *rc* points are considered weak representative cells and are discarded from the group of points in the last component. In this component, a $p_r$ is computed for each group. These representative $p_r$'s together form the *RT* (Algorithm 1, lines 12 to 14). To generate a $p_r$, the three MAT dimensions for all points in the cell, we summarize the three MAT dimensions for all points in the cell: spatial, temporal, and semantic.

For *spatial dimension*, the *centroid* point is computed (WOOD et al., 1990), i.e., the average of the (x,y) coordinates in *eachGroupPoint*. For the *temporal dimension*, we compute the *Significant Temporal Intervals (STI)* that accommodate all timestamps within the set of points *eachGroupPoint*. The STI captures the time intervals during which these points are distributed.

**Definition 4.1.1.** ***Significant Temporal Intervals (STI).*** Let $Time = ts_1, ts_2, \ldots, ts_m$, where each $ts$ represents the temporal value of each point. An STI is a collection of time intervals $[ts_{i+1} - ts_i], \ldots, [ts_{m-1} - ts_m]$ that encompasses all the $tsi$ values within the points of each cell grid ($tsi \in eachGroupPoint$). In this context, $i$ denotes the initial index in the *Time* sequence, and $m$ represents the final index.

To identify the most relevant $STI$s for the task of creating representative points ($p_r$), MAT-SG establishes a ranking based on the intervals ($sti \in STI$) and their corresponding temporal tendencies. We use the predefined threshold $\tau_{rv}$ to define which $sti$ are considered representative for $p_r$. Specifically, $sti$ intervals with a frequency rate greater than or equal to $\tau_{rv}$ are considered representative.

To illustrate this process, Algorithm 3 outlines the computation of the ranking for representative $STI$s. This process is exemplified in a visual representation in Figure 8.

First, a *Time* list is generated to hold all $ts$'s $\in$ *eachGroupPoint*. It is sorted for better analyzing the time intervals (lines 1 to 3), as shown in Figure 8 (a). Then, we consider a com-

**Algorithm 3:** MAT-SG:*computeTemporalDimension*

```
   input  : eachGroupPoint, τ_rv
   output: rankSTI                          /* ranking of representative STIs for eachGroupPoint */
1  foreach p ∈ eachGroupPoint do
2      Time.add(p.time);
3  Time.sort();
4  Δ_Time ← computeTimesDifference(Time);
5  V_ΔTime ← computeValidValues(Δ_Time);
6  τ_t ← computeTimeThreshold(V_ΔTime);
7  STI_aux ← ∅;
8  rankSTI ← ∅;
9  foreach ts ∈ Time do
10     STI_aux.append(ts);
11     if δ_i > τ_t and (|STI_aux|/|Time|) ≥ τ_rv then
12         rankSTI.new(STI_aux);
13         rankSTI.get(STI_aux).put(|STI_aux|/|Time|);
14         STI_aux ← ∅;
15 rankSTI ← normalizeRank(rankSTI());
16 return rankSTI
```

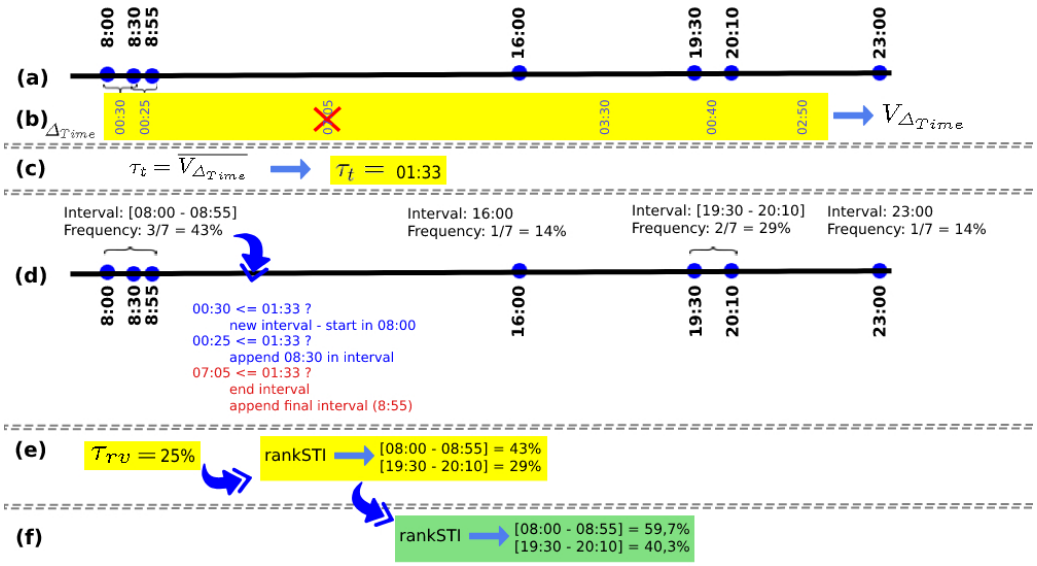


Figure 8 – An example of temporal dimension summarization in a grid cell

puted threshold ($\tau_t$) to define when a $ts \in Time$ is close to another and aggregate $ts$'s to generate an $sti$, as explained in the following.

Consider $\delta_i$ as the time difference between two consecutive timestamps ($\delta_i = ts_{i+1} - ts_i$), and let $\Delta_{Time} = \{\delta_1; \delta_2; ...; \delta_{n-1}\}$ represent a set of $\delta_i$ values for all $ts \in Time$ (line 4). It is important to note that, according to our conceptual model (presented in Figure 5), each point in the input dataset contains one Temporal Aspect, which could be either a single timestamp or a temporal interval defined by timestamps for the start and end times. In the latter case, the time difference is computed as two separate $ts \in Time$.

In line 5, we set the *Valid Temporal Interval set* $V_{\Delta_{Time}}$ as all $\delta_i \in \Delta_{Time}$ that fit into the average $\overline{\Delta_{Time}}$ plus or minus the standard deviation $\sigma\Delta_{Time}$, as defined by Equation 4.1.

$$V_{\Delta_{Time}} = \left\{\delta_i \in \Delta_{Time},\ 1 \leq i \leq (n-1)\ \mid\ \left(\overline{\Delta_{Time}} - \sigma\Delta_{Time}\right) \leq \delta_i \leq \left(\overline{\Delta_{Time}} + \sigma\Delta_{Time}\right)\right\} \tag{4.1}$$

To establish $STI$ that accurately represents the underlying patterns within each group

of points (*eachGroupPoint*), it is imperative to create a robust definition of these intervals adaptable to various scenarios. This adaptability ensures that the methodology remains effective across different datasets. To achieve this, we utilize a computational procedure to calculate $\tau_t$, aiming to define which are the *sti*'s in *eachGroupPoint*, by identifying when a $ts \in Time$ is in proximity to another and combining them to produce an $sti$. This procedure involves the identification and removal of potential outliers from the analysis.

In line 6, $\tau_t$ is determined as the average of the set $V_{\Delta_{Time}}$ ($\overline{V_{\Delta_{Time}}}$). Its purpose is to filter out $\delta_i \in \Delta_{Time}$ that differ significantly from the general trend, effectively removing outliers from consideration. This is illustrated in Figure 8 (b), where all $\delta_i \in \Delta_{Time}$ are shown, along with $V_{\Delta_{Time}}$. Notably, in this example, 07:05 is identified as an outlier. The subsequent Figure 8 (c) demonstrates the computation of $\tau_t$.

The construction of $STI$ is based on the calculated $\tau_t$ and is carried out in lines 9 to 14. We initially append to $STI_{aux}$ the values of $ts \in Time$ (line 10), and while $\delta_i$ is less than $\tau_t$, we consider $ts$ part of a $sti$ and continue to append subsequent timestamps. When $\delta_i$ exceeds $\tau_t$, and the frequency rate of this interval is considered representative (line 11), it is identified as a distinct $sti \in STI$. This $STI_{aux}$ is then added as a new key to the inverted list of $rankSTI$ (line 12), with its frequency rate serving as the value associated with this key (line 13). It is worth noting that a temporal interval $sti$ may also represent a single timestamp when it is considerably distant from its neighbors, meaning that the time differences $\delta_i$ to its adjacent points, regarding temporal information, exceed $\tau_t$. This process is illustrated in Figure 8 (d).

In the example, we have a first $sti_1$ = {08:00, 08:30, 08:55} as all their $\delta_i \leq \tau_t$. A $sti_2$ = {16:00} holds a single $sti$ as the time differences to its neighbors exceed $\tau_t$. This process is repeated to all the remaining $ts \in Time$. In Figure 8 (e), with a specified $\tau_{rv}$ of 25%, the identified $STI$s are $\{[08:00-08:55],[19:30-20:10]\}$. In the final step of the temporal dimension computation, performed in line 15, the resulting $STI$s are normalized to ensure that their rate values sum to 100%. This normalization process is depicted in Figure 8 (f).

We summarize the semantic dimension in the final step of $p_r$ Computation. This dimension encompasses various aspects, which we categorize into two types: *(i)* categorical, such as the mean of transportation and weather conditions, and *(ii)* numerical, such as air temperature and humidity. For numerical types, we calculate the *median* value[16].

We determine the representative mode values for categorical aspects, similar to the temporal dimension. These mode values appear most frequently within each aspect of the data grid cell and are identified based on a predefined threshold ($\tau_{rv}$). Once identified, we normalize the values to ensure the proportion values add up to 100%.

To illustrate this process with a practical example, consider a group of ten data points associated with activities performed in each place. Among these points, four are labeled as "tourism", four as "work", and two as "study". Initially, when applying MAT-SG, the mode values are "tourism" and "work", with each representing 40% of the data, while "study" accounts

[16] We prefer the median value instead of the mean value when the data are not symmetrically distributed since it is less sensitive to the influence of outliers (MCCLUSKEY; LALKHEN, 2007).

for 20%. However, with a representative value threshold set at $\tau_{rv}$ = 25%, the "study" value falls short of meeting the threshold and is consequently excluded as a representative value. In this case, "tourism" and "work" are considered representative values, and their proportions are adjusted to reflect the distribution of the most common activities in the dataset, with each now representing 50% of the representative values. This reorganization ensures that the representation accurately reflects the distribution of the most common activities in the dataset, providing an informative summary of the categorical data.

Overall, the $p_r$ computation step combines the computation of centroids, *sti*'s, and representative values for numerical and categorical aspects. This step consolidates the summarized information for each dimension, essential in determining the *RT*.

#### *4.1.3.3 Computation of the Better Representative Trajectory*

To analyze and compute the better *RT* (according to Figure 6 step 7), MAT-SG employs a representativeness measure called RMMAT (details provided in Chapter 5). This measure is based on a similarity measure and the covered MAT points.

The representativeness measure is computed using the RMMAT function (Algorithm 1, line 15). This function calculates the representativeness measure between the input MATs (**T**) and the computed *RT*. The chosen similarity measure for this implementation is *MUITAS* (PETRY et al., 2019), recognized as a state-of-the-art similarity measure for MATs. MUITAS quantifies the distance between points in two MATs to determine their similarity.

The RMMAT measure reflects the overall coverage of both MAT points and the information in the *RT*. To ensure equal consideration of both similarity and covered information, we employ a strategy with equal weights, setting $\omega_{sim} = \omega_{cover} = \frac{1}{2}$. The measure combines the similarity measure and coverage proportion, aiming to identify the *RT* that achieves the maximum coverage of both MAT points and their contained information.

In the MAT-SG method, spatial segmentation takes priority over other dimensions. This means that even in scenarios where all points within the same cell exhibit temporal and semantic differences, MAT-SG computes at least one representative point considering the spatial dimension. This prioritization emphasizes the representativeness of specific locations in the input MATs, thus ensuring that spatial information is adequately preserved in the *RT*.

## 4.2 MAT-SGT: MULTIPLE ASPECT TRAJECTORY SUMMARIZATION BASED ON A SPATIAL GRID AND TEMPORAL SEQUENCE

MAT-SG stands out as a pioneering approach to generating representative data tailored specifically for MATs. It accomplishes this by addressing all aspects of MATs individually, which leads to a more comprehensive representation of the data. Moreover, MAT-SG introduces a pivotal task by establishing a mapping between the input MATs and the resulting representa-

tive MAT. This mapping is instrumental in preserving the relationship between the original data and its summarized representation.

It is worth emphasizing that the MAT-SG method involves spatial segmentation and data summarization. This process is valuable for identifying movement patterns that are specific to different spatial areas. Additionally, it comprehensively considers various dimensions and treats each semantic type individually, which is beneficial for capturing the full spectrum of MATs.

However, the effectiveness of trajectory data summarization should be viewed in the context of the intended purpose of the representative data. In some scenarios, temporal information is critical for understanding when and how events or movements occur over time. While MAT-SG excels in various aspects of summarization, it may not fully capture the temporal dimension of the data. This limitation might be a crucial factor for certain applications or analyses that heavily rely on temporal patterns within MATs.

In light of the importance of considering the intended purpose of representative data and the significance of temporal information in some scenarios, a novel method for summarizing MATs has been introduced, called *Multiple Aspect Trajectory Summarization based on a Spatial Grid and Temporal Sequence* (MAT-SGT). This method is designed to address the limitation of not fully capturing the temporal dimension of the data while still maintaining some of the advantages of MAT-SG.

Similar to MAT-SG, MAT-SGT aims to reduce the input dataset while providing representative data encapsulating the predominant patterns within MATs. However, MAT-SGT takes a novel approach by specifically focusing on identifying the temporal sequences associated with movement patterns. This is a crucial addition, as it ensures that the summarization method can better reveal when and how events or movements occur over time within MATs.

In addition to capturing temporal sequences, MAT-SGT retains some key features from MAT-SG, such as establishing mappings between input MATs and the representative MAT, as well as incorporating spatial segmentation. This comprehensive approach allows MAT-SGT to consider both spatial and temporal aspects in MAT summarization.

Analyzing and extracting meaningful insights from MAT data, which includes spatial, temporal, and semantic aspects, can be challenging. Considering this issue, our method analyzes the distribution of points over time and space to identify information values that best represent the main behavior exhibited in the input MATs. By leveraging spatiotemporal analysis techniques, we can capture patterns in movement, providing valuable insights into the overall trajectory data with a focus on the spatiotemporal sequence.

### 4.2.1 Data model

In maintaining the representative MATs computed by MAT-SGT a conceptual data model, illustrated in Figure 9, is employed.

This model provides a standardized representation of the input data and preserves the

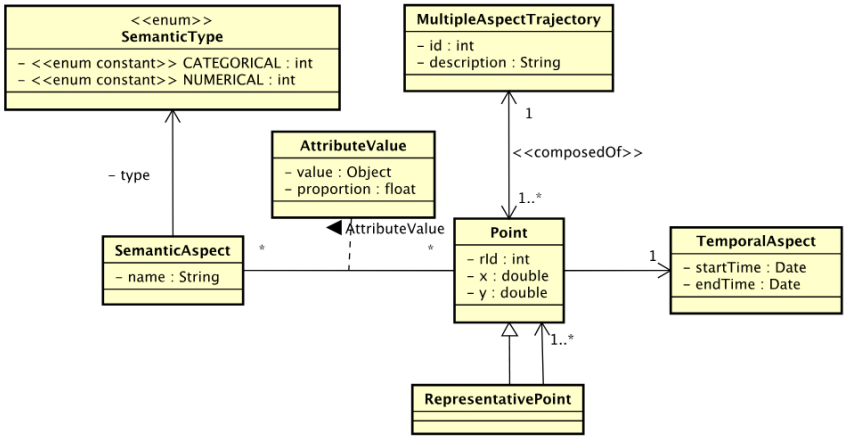


Figure 9 – The conceptual model for MAT-SGT

relationships between representative points and their corresponding input points, similar to the approach in MAT-SG. Each point in this model contains information related to spatial, temporal, and semantic dimensions. The semantic dimension consists of a set of aspects along with their respective values. The *RT* (representative MAT) is organized as a set of $p_r$'s that effectively summarizes the information from the MAT. Each $p_r$ summarizes relevant data derived from input MAT points, and a robust relationship is maintained between each $p_r$ and its associated MAT points to ensure the accuracy of the representation.

A notable distinction between MAT-SGT and MAT-SG lies in how they handle temporal aspects related to representative points. In the MAT-SG method, a $p_r$ could potentially encompass a set of Temporal Aspects. In contrast, MAT-SGT takes a different approach, where each $p_r$ is associated with a single Temporal Aspect, which could be a single occurrence or a temporal interval, maintaining only the relationship regarding the Point entity, as each $p_r$ is a specialized MAT point that preserves specific attributes. In essence, *RT* in MAT-SGT is constructed as a temporal sequence of $p_r$'s, highlighting the importance of temporal information in this method. This emphasis on temporal sequences is a distinctive feature that sets MAT-SGT apart from MAT-SG and enhances its capabilities for MAT summarization, especially in capturing the temporal information of mobility activity.

### 4.2.2 Architecture

Figure 10 provides an overview of the MAT-SGT method, comprised of two core components: *(i) Data Segmentation* and *(ii)* $p_r$ *computation*. The main objective of Data Segmentation is to reveal underlying data patterns focused on data density in both spatial and temporal dimensions. In contrast, the $p_r$ Computation component summarizes data by analyzing attribute value frequency.

The method receives as input a set of filtered MATs (**T**) based on specific criteria (step 1). Subsequently, the input MAT points are segmented into a spatial cell grid (step 2) to identify relevant cells. For each relevant cell, steps 4 to 6 are performed to compute representative points

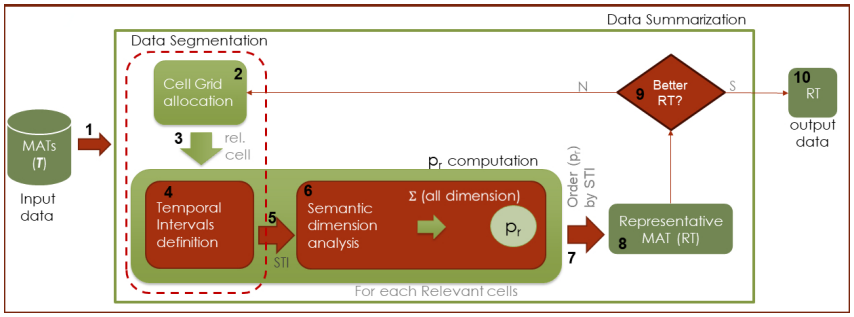


Figure 10 – Overview of the MAT-SGT method.

$p_r$ that summarize all dimensions and capture the essential characteristics of the input data within each cell.

All computed $p_r$'s are ordered based on the temporal dimension (step 7), resulting in the *RT* as the output data (step 8). The best of the computed *RT*'s is then selected as the final result (step 9). MAT-SGT provides a comprehensive representation of the main behaviors and characteristics exhibited by the input MATs, considering the spatial and temporal density as well as the frequency of each aspect attribute value. The next section details the MAT-SGT process.

Notably, MAT-SGT focuses on temporal summarization and in-depth analysis of other aspect attribute values. The temporal intervals defined in the *Data Segmentation component* play a pivotal role in this analysis.

In essence, MAT-SGT builds upon the MAT-SG methodology by incorporating temporal interval definition and temporal summarization. The overarching goal is to provide a richer and more comprehensive representation of input MATs, taking into account spatial and temporal density, the frequency of aspect attribute values, and a deeper understanding of temporal aspects in the data analysis process.

### 4.2.3 Algorithm

MAT-SGT algorithm considers the same input parameters as MAT-SG, as detailed in Table 4 (Algorithm 4). Like MAT-SG, it first computes the minimum spatial threshold ($\tau_s$) 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 that is furthest away from it (line 5). Since the initial grid is based on the initial $z$ value (lines 9 and 10), this cell size is iteratively reduced, aiming to compute a better *RT* (lines 8 to 26).

Both MAT-SG and MAT-SGT algorithms aim to find the optimal segmentation for a better *RT*. One of the main differences in MAT-SGT is in line 11, where it aims to find the optimal segmentation for a better *RT*. Each iteration segments data spatiotemporally, based on the current $z$ value, providing spatial allocation (*Cell Grid allocation* step), and calculates representative points by analyzing the temporal intervals for each group of points. The second main difference in this method refers to the temporal sequence of representative points that

generate the *RT* (line 15).

The MAT-SGT algorithm accomplishes MAT summarization through two internal components: *(i) data segmentation*; and *(ii) $p_r$ computation*. The quality of the resulting *RT* is compared to the previous (*betterRT*). If it improves by at least 10%, *betterRT* gets updated (lines 17 to 21). The algorithm stops and returns the best *RT* if no improvements are found in two iterations. The two components of the MAT-SGT method are detailed next.

**Algorithm 4:** MAT-SGT

```
   input : T, τ_rc, τ_rv
   output: RT                                          /* representative trajectory */
 1 rc ← |T.points| × τ_rc;
 2 τ_s ← computeMinSpatialThreshold();
 3 rt, betterRT ← ∅
 4 betterRTmeasure, count ← 0;
 5 z ← computeMaxZValue();
 6 w_sim ← 0.5;
 7 w_cover ← (1 − w_sim);
 8 while z > 1 do
       // component (i) - Fig. 10 (steps 2 and 3)
 9    cellSize ← computeCellSize(τ_s, z);
10    relCells ← cellGridAllocation(rc, cellSize);
       // components (i) and (ii) - Fig. 10 (step 4 and 5)
11    setGroupPoints ← STIdefinition(relCells, τ_rv);
       // component (ii) - Fig. 10 (step 6)
12    foreach eachGroupPoint ∈ setGroupPoints do
13       p_r ← computeRepPoint(eachGroupPoint, τ_rv);
14       rt ← rt ∪ p_r
15    rt.sort(); // order by STI - Fig. 10 (step 7)
       // analysis of better RT - Fig. 10 (step 9)
16    rtMeasure ← RMMAT(rt, T, w_sim, w_cover);
17    if (rtMeasure × 1.1) ≥ betterRTmeasure then
18       betterRTmeasure ← rtMeasure;
19       betterRT ← rt;
20       rt ← ∅
21       count ← 0;
22    else
23       count++;
24    if count > 1 then
25       break;
26    z ← z × 0.85;
27 return betterRT;
```

#### *4.2.3.1 Data Segmentation Component*

This component performs data segmentation in two steps: *(i) Cell Grid Allocation* and *(ii) Temporal Intervals Definition*. In the first step, the cell size is computed based on the value of $z$ and $\tau_s$. This cell size determines the granularity of the spatial segmentation. Next, it allocates the input MAT points into the corresponding cells of the spatial grid. This step is similar to the spatial segmentation step of our previous method (MAT-SG). After allocating points, the method identifies *relevant cells* with at least *rc* points for insights.

In the second step, MAT-SGT analyzes the relevant cells to compute *Significant Temporal Intervals (STI)*. This step is similar to the *computeTemporalDimension* algorithm of MAT-SG (Algorithm 3). However, MAT-SGT introduces an additional refinement, where this analy-

sis serves both the purpose of data segmentation and the computation of representative points. For data segmentation, the STI rank is computed for each relevant cell. It involves computing and analyzing all temporal intervals within the cell and their tendency, determining which intervals can be considered representative based on a frequency rate threshold of $\tau_{rv}$. By applying this procedure, MAT-SGT defines the STI within each relevant cell, capturing the temporal patterns and characteristics of the input MATs. Then, it groups MAT points, each group defined by each $sti \in STI$ of its corresponding relevant cell (Algorithm 4, line 11). This grouping allows for the identification and extraction of meaningful points that share similar temporal characteristics.

#### *4.2.3.2 Representative Point Computation Component*

The second component of MAT-SGT summarizes the groups of points obtained from the initial component. This entails the computation of a representative point ($p_r$) for each group, and these $p_r$'s are sorted into a temporal sequence, ultimately forming the $RT$. The $p_r$ generation process comprehensively addresses the spatial, temporal, and semantic dimensions.

For the spatial dimension, the algorithm calculates the centroid of the points within each group. In the temporal dimension, we utilize the $sti$ as previously explained. Different strategies are applied when dealing with semantic dimensions, which can include both *categorical* and *numerical* aspects.

For numerical attributes, such as temperature or air humidity, MAT-SGT computes the median value as the representative value. In contrast, categorical attributes like transportation means or weather conditions rank the representative mode values. The mode signifies the most frequently occurring value for each aspect within the group. To determine which values are considered representative, a predefined threshold ($\tau_{rv}$) is applied, similar to our previous method. After identifying the representative values, these values are normalized to ensure that they collectively sum to 100%, effectively representing the distribution of these values within the group. This normalization ensures an accurate reflection of the categorical data distribution, delivering an informative summary.

In summary, the $p_r$ computation step combines centroid computation, utilization of $sti$, and representative value determination for both numerical and categorical aspects. This comprehensive approach consolidates the summarized information for each dimension, contributing to the computation of the $RT$.

#### *4.2.3.3 Computation of the Better Representative Trajectory*

To analyze and compute the better $RT$ (according to Figure 10 step 9), MAT-SGT employs a representativeness measure called RMMAT (details provided in Chapter 5) that reflects the overall coverage of both MAT points and the information in the $RT$. This measure is computed in line 16 in Algorithm 4. This analysis sets $w_{sim}$ and $w_{cover}$ to equal values. The

measure combines the similarity measure and coverage proportion, aiming to identify the *RT* that achieves the maximum coverage of both MAT points and their contained information.

In the MAT-SGT method, spatiotemporal segmentation takes priority over other dimensions. It 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 that considers spatial and temporal dimensions. This approach highlights 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 specific moment, taking into account the dynamic nature of the data.

## 4.3 OUTPUT DATA

Both MAT-SG and MAT-SGT compute a representative MAT (represented by *RT*) which is outputted as a CSV file. The structure of the CSV file is determined by: (i) the configuration settings for the *RT* computation, and (ii) the information of each representative MAT point. The configuration settings include: $CellSize$, $\tau_{rc}$, $\tau_{rv}$, |cell|, minPointRC, $|RT|$, and $|coverPoints|$. Here is a breakdown of what each setting represents:

- $CellSize$ 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 has the following structure: "$lat\_lon$, $time$, $\#Semantic\_Aspects\#$, $mapping$". The "$lat\_lon$" refers to the spatial dimensions of the point made up of latitude and longitude. The "$time$" refers to the temporal aspects of the point, which can be either an interval or a single occurrence. In MAT-SG, the "$time$" information can be a rank, whereas in MAT-SGT, each $p_r$ is represented by only one-time value. The "$\#Semantic\_Aspects\#$" illustrated all the semantic aspects of the input MATs. These are categorical types that provide a normalized rank of information. For example, weather conditions can be ranked as follows: "{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," the present $p_r$ is composed of points with ID #3 and #9 of the trajectory ID #127, along with other points.

## 4.4 RUNNING EXAMPLE

This section provides an illustrative example of both methods, MAT-SG and MAT-SGT in order to highlight their differences. We consider a set of input MATs, denoted as $\mathbf{T} = \langle q, r, s \rangle$. Each MAT, represented as $q = \langle p_{q_1}, p_{q_2}, ..., p_{q_n} \rangle$, $r = \langle p_{r_1}, p_{r_2}, ..., p_{r_m} \rangle$, and $s = \langle p_{s_1}, p_{s_2}, ..., p_{s_t} \rangle$, corresponds to the trajectory of a same individual in the different days. Figure 11 presents these MATs along with related aspects such as the *price* spent at PoIs, the visited *PoIs*, *weather conditions*, and *rain precipitation*.

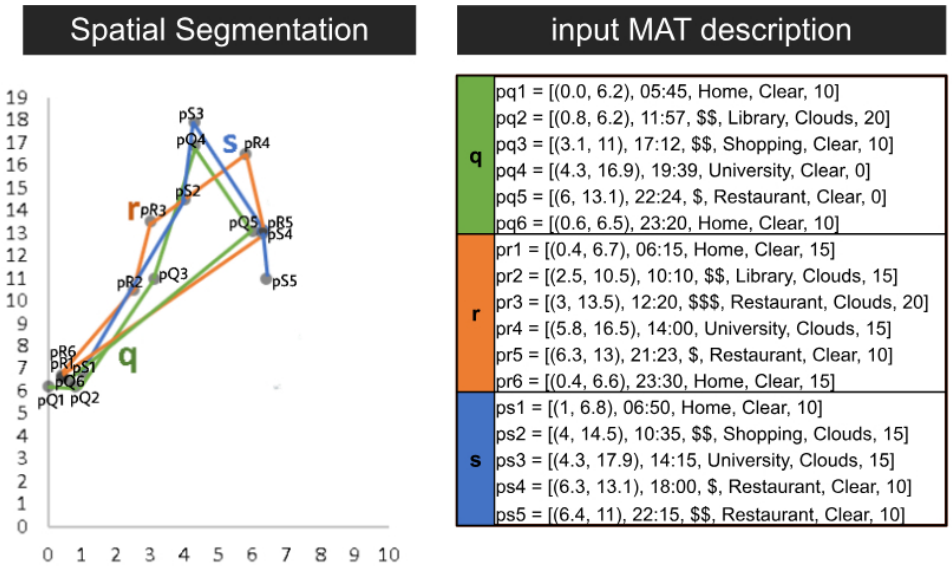


Figure 11 – Sample data with point aspects information for trajectories $q$, $r$, and $s$.

In this example, we set the input values as $\tau_{rc}$= 25% and $\tau_{rv}$= 25%. Given that the total number of points in $\mathbf{T}$ is 17, a relevant cell must contain more than 4 points. Additionally, we define a cell size of 12.5 to ensure a consistent structure and facilitate understanding of the summarization process in both methods.

### 4.4.1 MAT-SG

Figure 12 presents the resulting representative trajectory $rt = \langle p_{rt_1}, p_{rt_2} \rangle$ from different perspectives. In Figure 12 (a), the spatial distribution of the representative trajectory computed from $\mathbf{T}$ is showcased. The input MATs are segmented into a grid of cells, and the red line denotes the corresponding $RT$. Figure 12 (b) provides a detailed output, offering additional information and insights about the $RT$. As previously mentioned, data summarization occurs within cells containing more than 4 points.

For a more in-depth understanding of our summarization process, each step is illustrated in Figure 13. Let's focus on the first cell, as shown in Figure 13 (a). In step (b), which corresponds to the $p_r$ *Computation* step, each dimension is summarized. The temporal and semantic dimensions are highlighted, considering $\tau_{rv}$= 25%, and the values considered representative for each aspect are identified, i.e., those with at least a $\tau_{rv}$ value. Subsequently, these

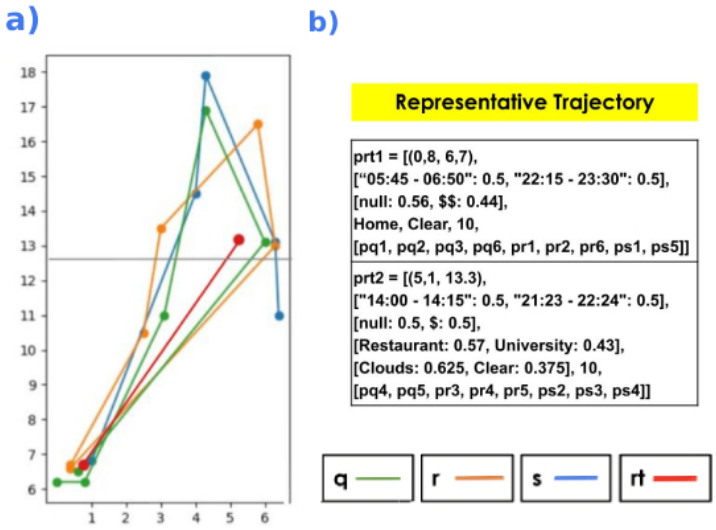


Figure 12 – Visualization of the resulting MAT-SG representative trajectory ($RT$) from different perspectives: (a) Spatial view; and (b) Detailed $RT$ description of point aspects, providing additional insights.

representative values for each aspect are normalized, and the resulting values across all aspects represent the $p_r$ of this cell. In this instance, $p_{rt_1}$ serves as the referent MAT point for the first cell, derived from $p_{q_1}$, $p_{q_2}$, $p_{q_3}$, $p_{q_6}$, $p_{r_1}$, $p_{r_2}$, $p_{r_6}$, $p_{s_1}$, and $p_{s_5}$, as illustrated in Figure 15 (c).

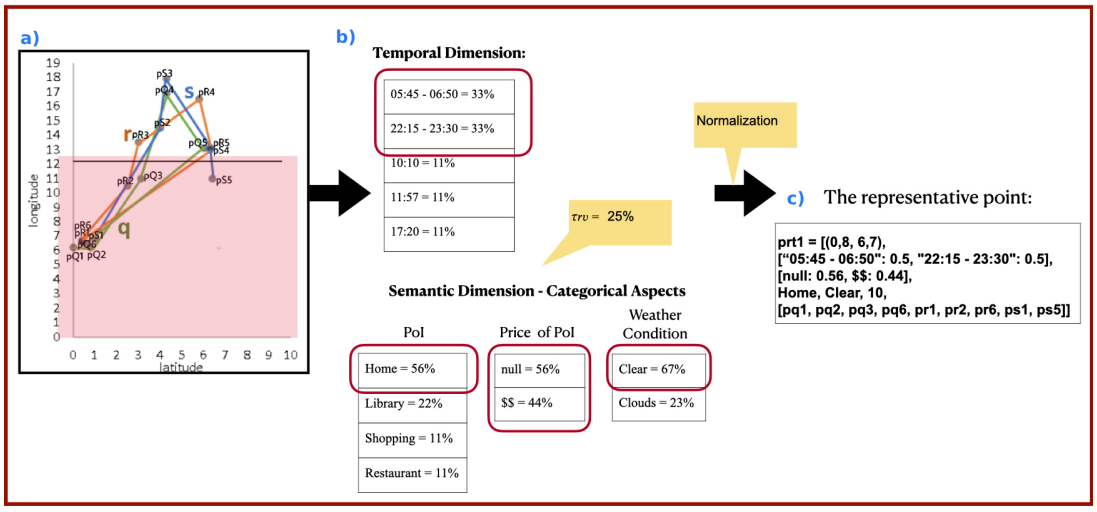


Figure 13 – A step-by-step perspective of the summarization process in MAT-SG, illustrated by the analyzed cell (a), the $p_r$ Computation step (b), and the final representative points computed (c).

In this scenario, insightful observations can be made, such as the individual's tendency to check in at home in the morning (between 05:45 and 05:50) and during the night period (between 22:15 and 23:30), likely corresponding to the times of leaving and returning to her/his residence.

### 4.4.2 MAT-SGT

Figure 14 shows the resulting $rt = \langle p_{rt_1}, p_{rt_2}, \ldots, p_{rt_k} \rangle$ in different perspectives. Figure 14 (a) shows the spatial distribution of the representative trajectory computed from **T**. The

input MATs are segmented into a grid of cells, and the red line indicates the corresponding *RT*. Figure 14 (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 14 (c), providing additional information and insights about the *RT*.

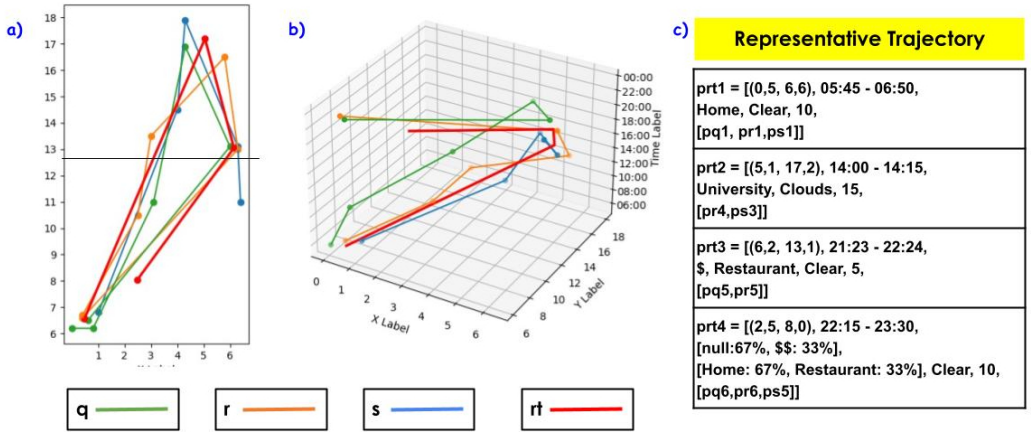


Figure 14 – Resulting MAT-SGT in representative trajectory (*RT*) visualization in different perspectives: (a) Spatial perspective; (b) Spatiotemporal perspective; and (c) *RT* description of point aspects providing additional details.

To gain a better understanding of our summarization process, we have illustrated each step in Figure 15. Let's focus on the first cell, as shown in Figure 15 (a). In step (b), which refers to the *Temporal Intervals definition* step, we identify 5 temporal intervals, consisting of 2 temporal intervals and 3 simple occurrences. Considering a $\tau_{rv}$= 25%, only 2 temporal intervals were considered as relevant $sti \in STI$ from the input MATs in this cell. The first *sti* covers the time interval between 05:45 and 05:50, while the second covers 22:15 to 23:30. These *sti*'s contain important MAT points that contribute to the computation of *RT* considering spatiotemporal density. Moving on to step (c) in Figure 15, it illustrates the $p_r$ *Computation* step, where the group of points in each relevant *sti* are summarized into a representative point. Specifically, $p_{rt_1}$ represents the referent MAT point for the first segment (derived from $p_{q_1}$, $p_{r_1}$, and $p_{s_1}$), and $p_{rt_4}$ represents the referent MAT point for the second segment, as illustrated in Figure 15 (d).

In this scenario, it is inferred that the individual typically leaves home between 05:45 and 06:50. After spending time during the day, he/she tends to have dinner near home, likely in a more affordable restaurant, before returning home.

## 4.5 SUMMARY

This chapter addresses the challenges of summarizing trajectories, particularly focusing on reducing trajectory data volume and preserving key patterns, as described on Section 2.3, with a specific emphasis on the complexities highlighted in MATs, such as data volume, velocity, and complexity (as described in Section 2.2). Our main contribution are 2 novel methods,

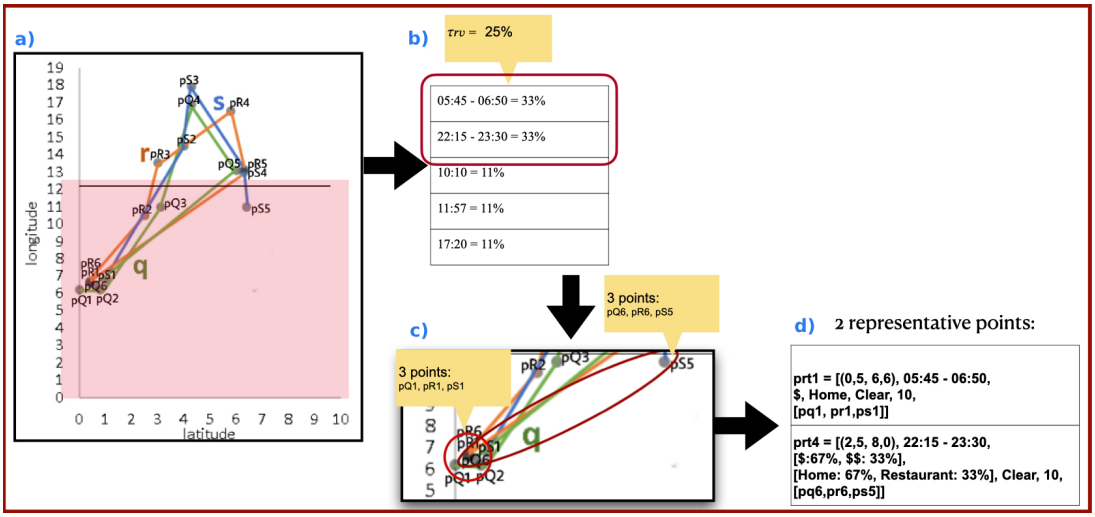


Figure 15 – A step-by-step perspective of the summarization process in MAT-SGT, illustrated by the analyzed cell (a), the Temporal Definition step (b), the $p_r$ Computation step (c), and the final representative points computed (d).

MAT-SG and MAT-SGT, designed to generate a representative MAT for a given set of MATs filtered by some criteria.

In essence, both MAT-SG and MAT-SGT methods aim to compute a representative MAT (*RT*) by delving into the distribution of MAT points. These approaches systematically identify and prioritize significant segments and aspects within the input MATs, resulting in an *RT* that comprehensively captures the main behaviors and characteristics of the input MATs. This ensures a succinct summary of each aspect individually. Moreover, both methods establish a coherent mapping between the input MATs and the resultant representative MAT, thereby preserving the intrinsic relationship between the original data and its summarized representation.

While MAT-SG specializes in spatial segmentation/density and data summarization, aiding in the identification of movement patterns across various spatial areas and addressing various aspects, i.e., MAT-SG excelling in summarizing representative aspects within specific spatial areas. This method is invaluable when understanding spatial patterns is crucial, regardless of temporal sequence relevance. For example, in the context of vessel trajectories, where the purpose is to identify the region where each activity happens (cargo or fishing) and the pattern aspects involved in each activity, MAT-SG works well.

In contrast, MAT-SGT focuses on emphasizing temporal sequences, providing detailed insights into the chronology of events or movements over time. It focuses on computing a representative MAT by identifying the temporal sequences associated with movement patterns. For instance, in the context of a recommendation system, where understanding the sequence movement over time, identifying the region, trend period, and aspects involved, MAT-SGT could be better suited.

Despite serving distinct purposes, both methods share the common goal of identifying and prioritizing significant segments and aspects, culminating in a representative MAT (*RT*) that effectively captures the pivotal features of the input MATs. However, the choice between

MAT-SG and MAT-SGT depends on the analysis objectives, with MAT-SG preferred for spatial pattern comprehension and MAT-SGT for detailed temporal narratives.

# 5 RMMAT: REPRESENTATIVE MEASURE FOR MULTIPLE ASPECT TRAJECTORIES

This chapter presents another contribution of this Thesis. It introduces the *Representativeness Measure for Multiple-Aspect Trajectories (RMMAT)*[17] , a standardized metric for evaluating the effectiveness of representative data given by summarization methods, offering a solution to the challenge of evaluating how effectively a representative trajectory reflects the original dataset (MACHADO et al., 2023b). RMMAT leverages similarity metrics and covered information to offer a comprehensive measure that quantifies the quality of representative data concerning the complete input dataset. This score can be customized to align with the specific requirements of different analytical scenarios, allowing analysts to tailor the evaluation process accordingly.

The core question addressed in this section is: 'How much of the *RT* captures and reflects the original MATs' essence within an input dataset **T** ($D = t_1, t_2, ..., t_n$ )?'. The computation of *RT* should be customizable based on specific use case objectives and requirements, as different applications may demand varying levels of granularity and information preservation.

Designed for big trajectory data with multiple aspects, this novel representativeness measure aims to quantify the information coverage of *RT* from the input dataset **T** and estimate its similarity to the entire dataset, i.e., 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. The objective is to simplify the evaluation of summarization methods and extract valuable insights from extensive MAT datasets.

RMMAT is designed to provide a balanced and objective measure of two components: (i) similarity metric and (ii) covered information. By assigning numerical values to similarity, this measure offers a concrete and measurable way to assess how closely the *RT* reflects the complex patterns in the input dataset. Additionally, the measure takes into account the covered information, allowing us to evaluate whether the *RT* can accurately encapsulate specific points from the dataset, thus reflecting the overall integrity of the *RT* concerning the entire dataset. By combining these two components, RMMAT aims to address the limitations of evaluating representativeness in summarized MAT, providing a rigorous and objective evaluation of how well the *RT* captures the intricacies of the data. Both components are detailed next.

## 5.1 SIMILARITY METRIC COMPONENT

Trajectory similarity serves as a metric to measure the similarity between two trajectories, considering the entire movement, involving attributes like spatial positions, temporal sequences, and semantic aspects. This method helps in assessing how much common patterns exist in the movement of two trajectories. While traditional similarity measures are effective for comparing individual trajectories, computing the similarity of a particular trajectory, the *RT*,

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

against all other trajectories in a dataset is still an open issue.

To address this issue, we evaluate the similarity measure between $RT$ and each trajectory $t_1, t_2, ..., t_n$ in $\mathbf{T}$, where both $\mathbf{T}$ and $RT$ are non-empty elements. Recognizing the potential presence of skewed distributions or outliers in the dataset, we decided to use the median value of the similarity measure across all pairs of MATs ($RT$ and each $t \in \mathbf{T}$), given that $0 \leq \text{Similarity} \leq 1$. By using the median, a more robust measure of central tendency than the average, we guard against the influence of outliers and skewed data distributions. It ensures that extreme values or anomalies in similarity scores do not unduly impact the result, providing a more balanced representation of central tendency. The equation is expressed as follows.

$$\begin{aligned} |Similarity(RT, \mathbf{T})| = \\ Me(\{Similarity(RT, t_1), Similarity(RT, t_2), ..., Similarity(RT, t_n)\}) \end{aligned} \tag{5.1}$$

The function $Me$ calculates the median similarity score between $RT$ and all $t \in \mathbf{T}$ by computing the median of the similarity measures.

## 5.2 COVERED INFORMATION COMPONENT

Aiming to evaluate the accuracy on which $RT$ encapsulates specific information from $\mathbf{T}$, the covered information within $\mathbf{T}$ by $RT$ is computed. So, the covered MAT points by $RT$ in each $t \in \mathbf{T}$ are computed, i.e., the total MAT points in $\mathbf{T}$ that are mapped contribute to the computation of $RT$. The resulting proportion represents the covered information, a non-negative value indicating the overall integrity of the $RT$ relative to the entire dataset. This computation is defined as:

$$T^c(RT) = \left( \frac{\sum_{p \in t}^{\mathbf{T}} p \subseteq RT}{|\mathbf{T}.points|} \right) \tag{5.2}$$

The mapping between the input MATs and the representative MAT allows determining how much the computed $RT$ covers the input MATs ($T^c(RT)$). Equation 5.2 calculates the proportion of covered MAT points by the $RT$ concerning all $t \in \mathbf{T}$, representing how well the computed $RT$ captures the points of the input MATs ($\mathbf{T}$.points).

RMMAT is designed to provide a representativeness measure score that balanced both components: (i) similarity metric and (ii) covered information, and it is calculated by the final equation RMMAT, where RMMAT $\in$ [0,1]:

$$RMMAT = \omega_{sim} \times |Similarity(RT, \mathbf{T})| + \omega_{cover} \times T^c(RT) \tag{5.3}$$

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

be adjusted based on the specific scenario to prioritize either the covered information ($\omega_{cover}$) or the similarity ($\omega_{sim}$).

## 5.3 RUNNING EXAMPLE

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 **T**, each one representing a trajectory attributed to a different individual.

For 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. Here, we opt to use the same Running Example presenting for MAT-SGT, disposed in Section 4.4.2, one state-of-the-art MAT summarization method, and the widely recognized MAT similarity measure *MUITAS* (PETRY et al., 2019). As exemplified in Chapter 4, both methods MAT-SG and MAT-SGT establish a mapping between the input data and the resultant representative trajectory, facilitating the inclusion of covered information in the computation of representativeness. We employ a balanced weights strategy by setting $\omega_{sim} = \omega_{cover} = \frac{1}{2}$.

In order 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. Regarding the summarization method, we will use the setup of MAT-SGT where $\tau_{rc}$= 0.1 and $\tau_{rv}$= 0.25.

The input MATs and their corresponding *RT* are shown in Figure 16. 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*.

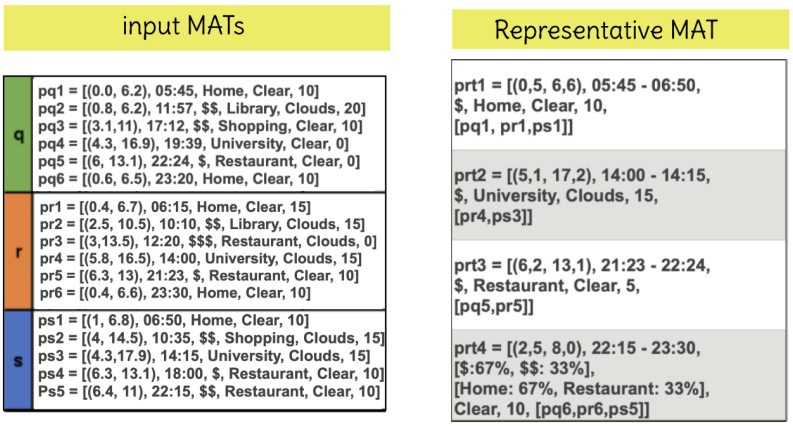


Figure 16 – Set of input MATs $\mathbf{T} = \langle q, r, s \rangle$, where $q = \langle p_{q_1}, p_{q_2}, ..., p_{q_n} \rangle$, $r = \langle p_{r_1}, p_{r_2}, ..., p_{r_m} \rangle$, and $s = \langle p_{s_1}, p_{s_2}, ..., p_{s_t} \rangle$ (left), and their correspondent *RT* (right).

For computing RMMAT, we first compute the similarity between each trajectory in

**T** and $RT$, where $MUITAS(q,RT) = 0.686$, $MUITAS(r,RT) = 0.835$, and $MUITAS(s,RT) = 0.871$. Then, according to Equation 5.1, the $|Similarity(RT,D)| = 0.835$. Regarding the covered information, Equation 5.2, $T^c(RT) = \frac{10}{17} = 0.5882$.

Finally, considering the computation of RMMAT with balanced weights strategy by setting $\omega_{sim} = \omega_{cover} = \frac{1}{2}$, and according to Equation 5.3, 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 **T** considering both similarity and covered information.

## 5.4 ANALYZING RMMAT REGARDING SIMILARITY INFORMATION

This section delves into the analysis of RMMAT focusing on similarity information. As (PETRY et al., 2019), we use the trajectories of each user as the ground truth, as trajectories of the same user are more likely to be similar than the trajectories of other users. Then, to gain insights into RMMAT behavior, we conducted an experiment using a sample of user trajectories of the Foursquare dataset (see Section 6.1).

Since there is no common strategy in the literature to evaluate a representative MAT for benchmarking, we established our criteria. For each group, we choose the MAT $t_i$ with the median similarity score as the baseline, computed across all group trajectories. It ensures that the baseline serves as a reference point for comparison purposes.

We present illustrative examples of evaluations based on the standard deviation (SD) of average and median similarity scores of each user's baseline. Three users were selected for in-depth analysis, each representing distinct characteristics in terms of SD: (i) user 185, showcasing a lower SD for average similarity scores; (ii) user 730, featuring a lower SD for median similarity scores; and (iii) user 708, displaying the highest SD for both average and median similarity scores.

This evaluation uses $\omega_{sim} = 1$ and $\omega_{cover} = 0$ based on the MUITAS similarity measure. The experiment involves assessing the representativeness of $RT$ in similarity information with different threshold values for relevant cell (RC) and representativeness value (RV), namely $\tau_{rc}$ and $\tau_{rv}$. The methods were repeated for each user with different parameter settings for $\tau_{rv}$ and $\tau_{rc}$, varying from 0% to 25% (0, 1, 5, 10, 15, 20, 25), to evaluate the sensitivity and robustness of the RMMAT measure. This investigation explores the impact of varying combinations of these thresholds on the computation of $RT$ in both MAT-SG and MAT-SGT.

On using MUITAS, we considered proximity functions, including spatial, temporal, and semantic functions, to assess the similarity between trajectories $T \in$ **T** and $RT$ with specified weights to balance all dimensions. The functions used are: *(i) spatial*: Euclidean distance measure. We consider a match occurs if the distance falls within a predefined threshold ($2 \times$ cellSize); *(ii) temporal*: we consider a match if the timestamp of $T$ falls within the temporal interval of $RT$; *(iii) semantic*: for numeric types, a match occurs if the difference is equal to or less than 10% of the $RT$ value, and for categorical types, a match occurs if the attribute value of $T$ falls within the range of $RT$ values.

Figures 17 and 18 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.

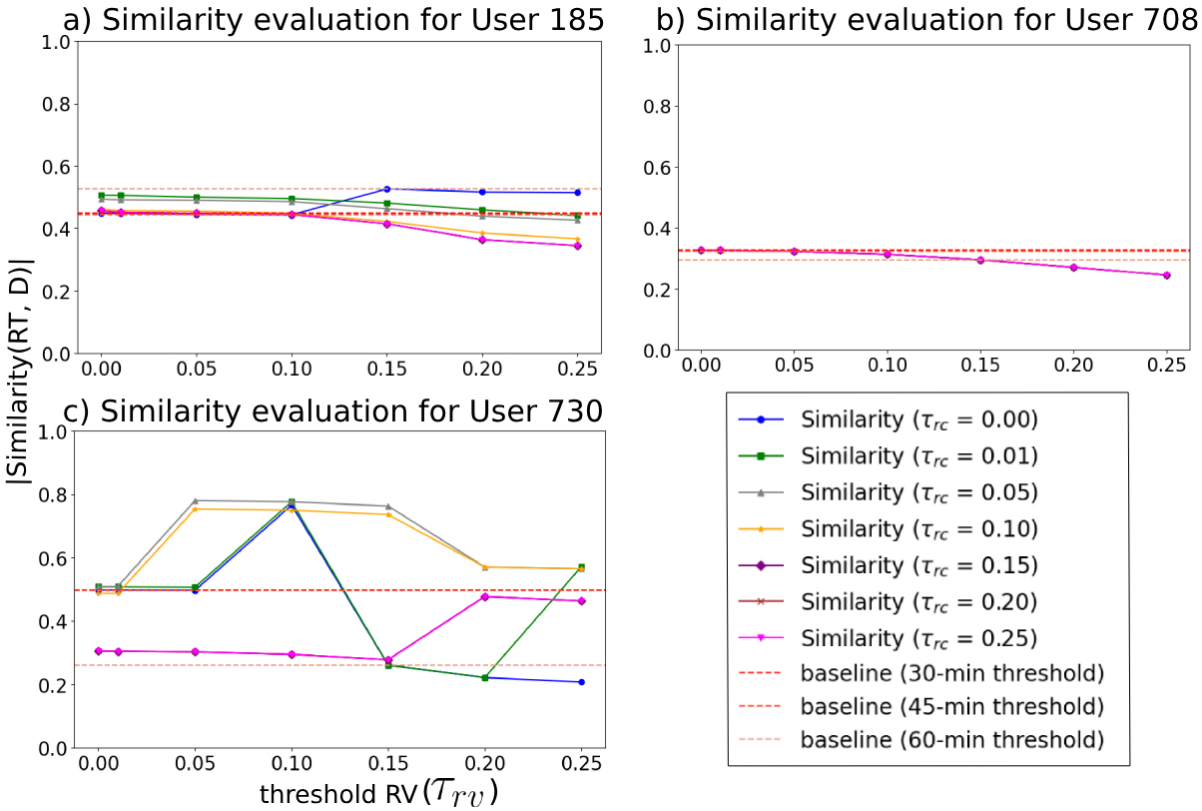


Figure 17 – This graph analyzes the similarity evaluation (Y-axis) by comparing varying threshold RC, the $\tau_{rc}$, shown as distinct lines, and the threshold RV, the $\tau_{rv}$, concerning baseline for users 185, 708, and 730. It explores different parameter configurations of the $\tau_{rv}$ (X-axis) to evaluate similarity. This analysis refers to the MAT-SG method.

Our RMMAT 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 RV threshold values. Regarding the threshold 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* computation.

For MAT-SGT, users 708 and 730 display specific *RT* behavior patterns across different RV threshold 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 threshold values for RC and RV in both methods on the RMMAT measure. The coefficients reveal relationships

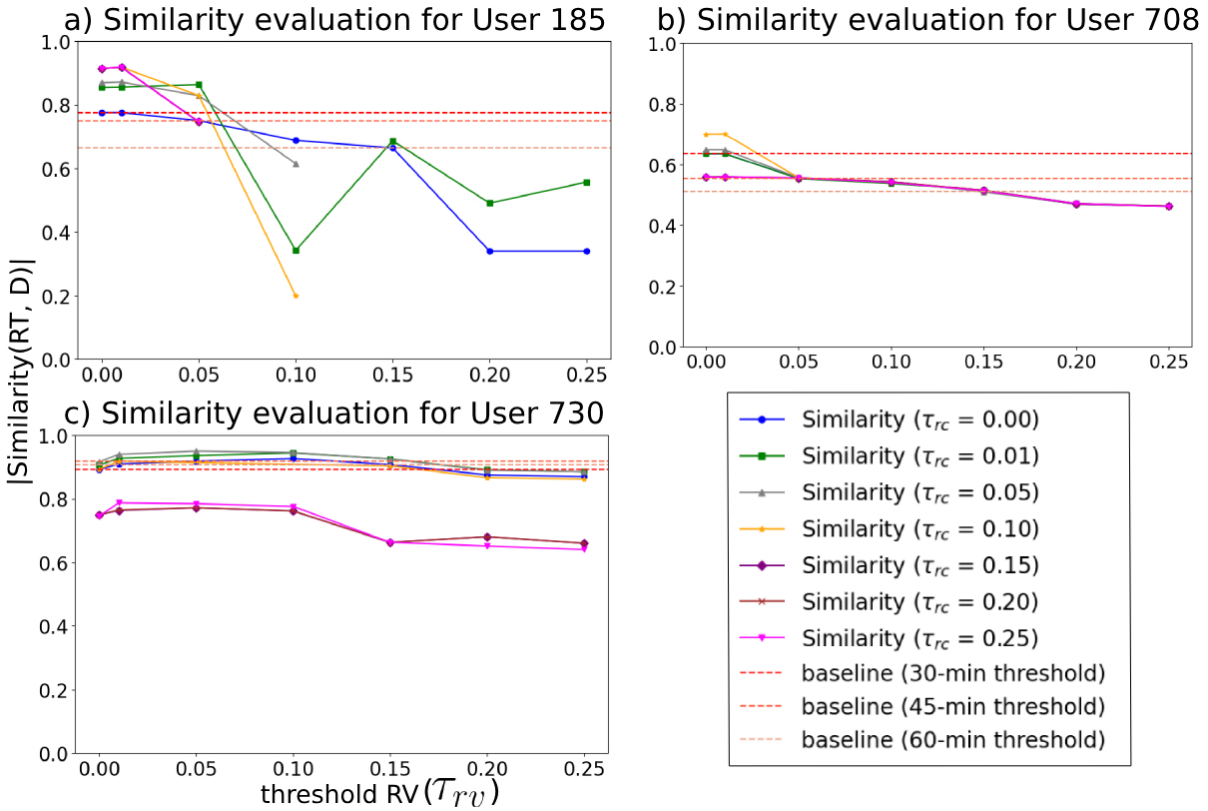


Figure 18 – This graph analyzes the similarity evaluation (Y-axis) by comparing varying threshold RC, the $\tau_{rc}$, shown as distinct lines, and the threshold RV, the $\tau_{rv}$, concerning baseline for users 185, 708, and 730. It explores different parameter configurations of the $\tau_{rv}$ (X-axis) to evaluate similarity. This analysis refers to the MAT-SGT method.

between input parameters and RMMAT scores for *RT* computed for both methods (MAT-SG and MAT-SGT) and input trajectories. The results in Table 5 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 5 – Impact of Input Parameters on the Representativeness Measure of *RT*

| **correlation coefficient** | **MAT-SG** | | **MAT-SGT** | |
|---|---|---|---|---|
| | **threshold RC** | **threshold RV** | **threshold RC** | **threshold 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 threshold RC, indicating that increasing threshold 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 threshold RC values consistently lead to lower RMMAT scores. For user 730, a negative correlation (-0.378) implies that higher threshold RC values result in lower RMMAT scores. Across all users in MAT-SG, the negative correlation pattern highlights that higher threshold RC values

lead to less representative *RT*.

For MAT-SGT, 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.

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 threshold RC values consistently lead to less representative *RT*. Meanwhile, in MAT-SGT, the correlation patterns reveal the nuanced impact of both threshold RC and RV values on RMMAT scores. The threshold RC and RV significantly influence 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.

## 5.5 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, specifically 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 $k = |T_{group}|$. 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). 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.

To assess the impact of covered information in RMMAT, we consider two scenarios for both MAT-SG and MAT-SGT regarding the computation of representativeness: (A) without covered information, which explores *RT* computation without considering covered information, and (B) with covered information, where covered information is integrated into *RT* computation. We obtain evaluation results by computing *RT* for each user in our selected sample in both MAT-SG and MAT-SGT with different threshold values for $\tau_{rc}$ and $\tau_{rv}$. These threshold values range from 0% to 25% (0, 1, 5, 10, 15, 20, 25), resulting in 49 runs for each user. By varying combinations of these thresholds, we explore their impact. We calculate similarity using MUITAS and order trajectories based on similarity scores. Finally, we employ the recall metric to measure the ability of *RT* to accurately rank trajectories within the same group. We highlighted the differences between both methods, emphasizing the higher value between with or without covered information.

Tables 6 (A) and (B) display the AR values for user 185 by MAT-SG in both scenarios, respectively. Additionally, Table 7 compiles the results of the AR analysis, where both scenarios yield the same outcomes for this specific situation.

Table 6 – The AR of User 185 by MAT-SG

(A) Without covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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) With covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

Table 7 – AR Analysis regarding covered information in User 185 by MAT-SG

| | **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 |

For the same user 185, the scenarios for MAT-SGT are respectively presented in Tables 8 (A) and (B), and the compiled results of the AR analysis are presented in Table 9. 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.

Upon analyzing the summarized outcomes of the AR analysis in Table 9, some relevant variations between including and excluding covered information for User 185 by MAT-SGT are observed. 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.

Table 8 – The AR of User 185 by MAT-SGT

(A) Without covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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) With covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 | - | - | - | - | - |

Table 9 – AR Analysis regarding covered information in User 185 by MAT-SGT

| | **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, computed by MAT-SG, Tables 10 (A) and (B) show the AR values, and Table 11 compiles the results of the AR analysis, where for this situation, both scenarios present the same results.

Table 10 – The AR of User 708 by MAT-SG

(A) Without covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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) With covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

Table 11 – AR Analysis regarding covered information in User 708 by MAT-SG

| | **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 |

By MAT-SGT, both scenarios for user 708 are respectively presented in Tables 12 (A) and (B), and the compiled results of the AR analysis are presented in Table 13. While there were

some minor variations in the specific values, the overall assessment presented in Table 13 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.

Table 12 – The AR of User 708 by MAT-SGT

(A) Without covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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) With covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 13 – AR Analysis regarding covered information in User 708 by MAT-SGT

| | **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 |

For the user 730, computed by MAT-SG, both scenarios are respectively presented in Tables 14 (A) and (B), and the compiled results of the AR analysis are presented in Table 15. In this situation, a slight variation can be observed 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.

Table 14 – The AR of User 730 by MAT-SG

(A) Without covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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) With covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

Table 15 – AR Analysis regarding covered information in User 730 by MAT-SG

| | **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 |

The AR values for user 730 computed by MAT-SGT in both scenarios are presented in Tables 16 (A) and (B). Additionally, Table 17 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 16 – The AR of User 730 by MAT-SGT

(A) Without covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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) With covered information

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 17 – AR Analysis regarding covered information in User 730 by MAT-SGT

| | **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 analysis of RMMAT w.r.t. covered information, as well as 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 point data. However, it is intriguing to observe that, for the same user, trajectories retrieved with covered point data fare better than computed RT trajectories, indicating a potential impact on RMMAT scores and implying differences in underlying data patterns.

In summary, 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 measure. It emphasizes the importance of considering each component in the RMMAT calculation to create a customized configuration that suits specific datasets and analysis objectives.

## 5.6 PROPERTIES OF RMMAT

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.

Addressing a critical gap, RMMAT introduces a quantitative metric for evaluating trajectory summarization methods. This objective measurement approach overcomes the limitations of previous subjective evaluation methods, paving the way for more accurate decision-making, deeper insights, and overall advancements in trajectory analysis.

The effectiveness of RMMAT in computing a representative MAT depends on the specific purpose and requirements of a use case. Different applications may need varying levels of granularity and information preservation. The evaluation of the summarization method is inherently tied to the specific objectives being analyzed. RMMAT focuses on assessing similarity and covered information, providing a comprehensive measure of the quality of representative data concerning the complete input dataset.

At last, RMMAT is not only a novel metric for trajectory data summarization. It also provides a flexible measure that can be adapted to diverse analytical scenarios. This adaptability, associated with the ability to objectively measure the quality of representative trajectories, makes RMMAT a valuable tool for researchers and analysts in the field of mobility analysis.

# 6 EXPERIMENTAL EVALUATION

As identified in Chapter 3, given the lack of a compatible baseline in related works, we opted to use datasets involving MATs for a quantitative evaluation. This chapter presents and details an experimental evaluation of both proposed summarization methods (MAT-SG and MAT-SGT), shedding light on their utility and representativeness. In the following sections, we describe the datasets, methodology, and the results of experiments over the proposed methods.

## 6.1 DATASETS

We evaluate the effectiveness of our methods using four datasets containing MATs. Three of these datasets, Foursquare, Gowalla, and Brightkite, are publicly available[18] . These datasets, widely employed in other works (ZHOU et al., 2018; PETRY et al., 2019; da SILVA; PETRY; BOGORNY, 2019; PORTELA; CARVALHO; BOGORNY, 2022), contribute to the robustness of our evaluation. Additionally, we include a private dataset[19] from Pisa, also utilized in Petry et al. (2019). 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; MYERS; LESKOVEC, 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.

The *Pisa dataset*, utilized in the evaluation of MUITAS (PETRY et al., 2019), was collected by 157 volunteers in Pisa through a mobile app, collected in Pisa, Italy, between May

[18] https://github.com/bigdata-ufsc/datasets_v1_0

[19] The TagMyDay experiment data was collected under a non-disclosure agreement during a visit funded by the SOBIGDATA Project in June 2023, so we cannot redistribute it. More information about it can be found at http://kdd.isti.cnr.it/project/tagmyday.

Table 18 – Datasets Vs attributes description

| Attribute | Type | Dataset | Description | Range / Example |
|---|---|---|---|---|
| Spatial | Numeric | All | Coordinates (latitude and longitude) | e.g.: lat: 40.83, long: -73.94 |
| Time | Temporal | All | Time of the day | [00:00,23:59] |
| PoI Category | Semantic: Categorical | Foursquare | The root type of category of PoI | e.g.: {Residence, Food, Transport,...} |
| PoI Price | Semantic: Categorical | Foursquare | The evaluation of Price | {-1,1,2,3,4} |
| PoI Rating | Semantic: Numerical | Foursquare | The rating evaluation of PoI | {-1} U [4.0, 10.0] |
| Weather | Semantic: Categorical | Foursquare | Weather condition | e.g.: {Clear, Clouds, Rain,...} |
| Weekday | Semantic: Categorical | All | Description of weekday | e.g.: {Sunday, Monday,...} |
| Time duration | Semantic: Numeric | Pisa | Time duration in Hours | {Up to 1h, 1 to 2h,...} |
| Trip purposes | Semantic: Categorical | Pisa | activity perform during the trip | e.g.: {Going home, Refueling, ...} |
| Transportation means | Semantic: Categorical | Pisa | transportation mode / type | e.g: {Car, Train, Bike, ...} |
| Distance traveled | Semantic: Numeric | Pisa | distance traveled in kilometers | {Up to 1 km,1 to 2 km,..., over 10 km} |
| PoI | Semantic: Categorical | Brightkite, Gowalla | the PoI definition | ID do PoI |

20, 2014 and September 30, 2014. It consists of movement segments representing users' daily routines, annotated with *transportation means*, *trip purposes*, *distance traveled*, *time duration*, and information of the *weekday*. After applying necessary transformations to ensure variability and consistency, including the removal of small trajectories with less than three segments and users with less than five trajectories, the final dataset comprises 9715 segments in 1617 daily trajectories from 92 different users. The trajectories exhibit an average length of approximately 26 trajectories per user.

Tables 18 and 19 provide detailed information about the attributes and characteristics of each dataset, respectively. Table 18 summarizes the datasets, including the attributes used in each dataset and their descriptions. Table 19 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 METHODOLOGY

The methodology section outlines the approach taken to evaluate the utility and effectiveness of the summarization methods (MAT-SG and MAT-SGT). We discuss the evaluation metrics employed to assess the performance of the methods, including similarity measures and the RMMAT metric. Additionally, we will detail the experimental setup, including parameter configuration and settings for key parameters such as $\tau_{rv}$ and $\tau_{rc}$, and their impact on the summarization results. We will also address other setups required to perform this experiment. By following a systematic methodology, we aim to provide a robust evaluation that captures the nuances of trajectory summarization.

### 6.2.1 Evaluation Metrics

Our experimental evaluation adopts a thoughtful and systematic approach to assess the utility of *RT*, employing two distinct strategies: (i) the Average Recall (AR) metric; and (ii) the

Table 19 – Summary of the used datasets

| Dataset | Description | Aspects |
|---|---|---|
| Foursquare | Traj Size: ~ 22<br># of Traj.: 3079<br># of Points: 66962<br># filtered data groups: 193<br>Filter Criteria: User | Lat, Lon, Time,<br>Weather Conditions,<br>PoI - Category, Price,<br>and Rating |
| Gowalla | Traj Size: ~ 18<br># of Traj.: 5329<br># of Points: 98158<br># filtered data groups: 300<br>Filter Criteria: User | Lat, Lon, Time,<br>PoI, Weekday |
| Brightkite | Traj Size: ~ 16<br># of Traj.: 7911<br># of Points: 130494<br># filtered data groups: 300<br>Filter Criteria: User | Lat, Lon, Time,<br>PoI, Weekday |
| Pisa | Traj Size: ~ 6<br># of Traj.: 1617<br># of Points: 9715<br># filtered data groups: 92<br>Filter Criteria: User | Lat, Lon, Time,<br>Time Duration, Weekday,<br>Transportation Means,<br>Trip Purposes,<br>Distance Traveled |

RMMAT.

The AR metric was inspired by the similarity measure work of Petry et al. (2019) and previously utilized in Section 5.5, which serves as our primary evaluation strategy. It helps us to evaluate the utility of *RT*'s within the context of the input dataset, thereby quantifying the quality of our summarization and representative data computation. AR measures the recall based on the similarity between the *RT* computed by each proposed method and other trajectories in the dataset.

Central to the AR metric is the computation of *RT* for each trajectory group filtered based on specific criteria. By dividing the dataset (**D**) into multiple groups ($T \in \mathbf{T} \in \mathbf{D}$) under the assumption that trajectories within the same group exhibit similarity, we aspire for high similarity values between the *RT* and trajectories within the same group. Indeed, we use the trajectories of each user as the ground truth for all datasets, as trajectories of the same user are more likely to be similar than the trajectories of other users.

The evaluation process unfolds systematically: *RT* is computed for each group, i.e., for each user in each dataset; a similarity search is conducted over the dataset; trajectories are ordered by similarity; and recall is calculated. The assessment hinges on the ideal scenario where the top $k$ most similar trajectories align with the same group trajectories ($k = |T_{group}|$). This metric effectively gauges the *RT*'s ability to rank trajectories within the same ground truth group.

In our second evaluation strategy, we turn our attention to measuring the representative-

ness of the representative MAT ($RT$) across the entire dataset. Our proposed representativeness measure facilitates this assessment, RMMAT, which aims to gauge the quality of $RT$ in terms of both similarity and covered information.

The RMMAT measure involves the computation of $RT$ for each group of filtered trajectories. The dataset ($\mathbf{D}$) is segmented into multiple groups ($T \in \mathbf{T} \in \mathbf{D}$). The RMMAT metric, ranging between 0 and 1, signifies the degree to which the $RT$ encapsulates the overall representativeness of the entire dataset. A value of 1 indicates that $RT$ fully represents the dataset, while a value of 0 implies that $RT$ fails to encompass any information from the dataset. To balance the consideration of both similarity and covered information, we adopt a strategy with equal weights, setting $\omega_{sim} = \omega_{cover} = \frac{1}{2}$. This ensures a comprehensive evaluation that accounts for both components of representativeness.

### 6.2.2 Experimental Setup

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). All experiments were implemented in Java and conducted on a Dell Inspiron laptop with an Intel Core i5 processor and 16 GB memory. The method was repeated on each user with a different setting of the parameters $\tau_{rv}$ and $\tau_{rc}$ with values varying from 5% to 25%, resulting in 25 runs for each user. We chose to start the parameter configuration at 5% because, for this analysis, it is not meaningful to consider lower information density when computing $RT$. The parameter $\tau_{rc}$ influences cell size, and consequently, the MAT points needed in each cell must be sufficiently dense and relevant. Similarly, extremely low tendency values imply that all values in the tendency will be considered representative. This parameter variation enables the evaluation of 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 are defined to assess the similarity between trajectories $T \in \mathbf{T}$ and $RT$, considering the distinct structure of $RT$. The adopted functions are *(i) spatial*: Euclidean distance measure. We consider a match if the distance between the spatial coordinates of the $T$ and $RT$ is within a predefined threshold ($2 \times$ cellSize); *(ii) temporal*: a match function based on the temporal interval of $RT$. We consider a match if the timestamp value of the $T$ falls within that interval; and *(iii) semantic*: functions for evaluating attribute matching for *numeric* and *categorical* types. We consider a match for numerical types if the difference between attribute values is equal to or less than 10% of the $RT$ value. For categorical types, we determine a match if the attribute value of the $T$ falls within the range of attribute values of the $RT$. W.r.t., for the weights parameter of MUITAS, we consider $w = 1/3$ for each dimension to balance all dimensions.

By adhering to established methodologies and introducing unique elements tailored to the goals of this study, these strategies form a robust foundation for the subsequent experimental evaluation, promising insightful findings into the performance of $RT$ in representing diverse

trajectory datasets.

## 6.3 RESULTS

In this section, we present the results of the evaluation of both the Average Recall (AR) metric for ranking user trajectories within the same group and the RMMAT as the representativeness measure, based on a specified parameter configuration, as previously described in the methodology.

### 6.3.1 AR Metric Strategy

The parameters $\tau_{rv}$ and $\tau_{rc}$ 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: (A) MAT-SG and (B) MAT-SGT.

#### *6.3.1.1 Foursquare-NYC dataset*

Table 20 displays the results for ranking user trajectories using AR. For MAT-SG, the highest value of 0.785 occurs with $\tau_{rv}$ and $\tau_{rc}$ both set to 0.05, while the lowest value (0.450) is obtained with $\tau_{rv}$ and $\tau_{rc}$ 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 20 – AR of ranking user trajectories in Foursquare dataset

(A) MAT-SG

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 | <u>0.450</u> |

(B) MAT-SGT

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 | <u>0.372</u> | 0.377 | 0.388 |

#### *6.3.1.2 Gowalla Location-Based Social Network dataset*

Table 21 provides the corresponding results for the Gowalla dataset. For MAT-SG, the highest value (0.871) is achieved with both $\tau_{rc}$ and $\tau_{rv}$ 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 $\tau_{rc}$ and $\tau_{rv}$ 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.

Table 21 – AR of ranking user trajectories in Gowalla dataset

(A) MAT-SG

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

(B) MAT-SGT

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

#### *6.3.1.3 Brightkite dataset*

Results for the Brightkite dataset are shown in Table 22. MAT-SG achieves the highest AR (0.928) with both $\tau_{rc}$ and $\tau_{rv}$ 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 $\tau_{rv}$ = 0.25 and $\tau_{rc}$ = 0.05. MAT-SGT showcasing superior performance under the best parameter configuration.

Table 22 – AR of ranking user trajectories in Brightkite dataset

(A) MAT-SG

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

(B) MAT-SGT

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

#### *6.3.1.4 Pisa dataset*

Table 23 displays results for the Pisa dataset. MAT-SG achieves the highest AR (0.687) with both $\tau_{rc}$ and $\tau_{rv}$ set to 0.05, while the lowest value (0.397) is identified with both parameters set to 0.25. MAT-SGT attains the highest AR (0.737) with $\tau_{rv}$ = 0.05 and $\tau_{rv}$ = 0.2, and the lowest value (0.487) is obtained with $\tau_{rv}$ = 0.25 and $\tau_{rc}$ = 0.05. MAT-SGT consistently demonstrates superior performance under the best parameter configuration.

Table 23 – AR of ranking user trajectories in Pisa dataset

(A) MAT-SG

| $\tau_{rv}$ \ $\tau_{rc}$ | **0.05** | **0.1** | **0.15** | **0.2** | **0.25** |
|---|---|---|---|---|---|
| **0.05** | **0.687** | 0.583 | 0.497 | 0.471 | 0.452 |
| **0.1** | 0.659 | 0.547 | 0.467 | 0.452 | 0.431 |
| **0.15** | 0.632 | 0.536 | 0.466 | 0.459 | 0.413 |
| **0.2** | 0.610 | 0.554 | 0.483 | 0.460 | 0.399 |
| **0.25** | 0.607 | 0.536 | 0.472 | 0.448 | 0.397 |

(B) MAT-SGT

| $\tau_{rv}$ \ $\tau_{rc}$ | **0.05** | **0.1** | **0.15** | **0.2** | **0.25** |
|---|---|---|---|---|---|
| **0.05** | **0.707** | 0.638 | 0.547 | 0.527 | 0.511 |
| **0.1** | 0.661 | 0.575 | 0.482 | 0.444 | 0.425 |
| **0.15** | 0.608 | 0.533 | 0.445 | 0.416 | 0.399 |
| **0.2** | 0.565 | 0.486 | 0.409 | 0.399 | 0.371 |
| **0.25** | 0.539 | 0.457 | 0.408 | 0.386 | 0.367 |

In general, MAT-SG exhibits a linear AR result when ranking user trajectories for each $\tau_{rc}$ across a range of $\tau_{rv}$. As $\tau_{rc}$ decreases, AR tends to decrease due to the algorithm's minimum requirement of MAT points in each cell for relevance. Conversely, MAT-SGT displays an inverse pattern, maintaining a linear AR result for each $\tau_{rv}$ across a range of $\tau_{rc}$, and as the value of $\tau_{rv}$ decreases, the AR also tends to decrease.

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 $\tau_{rc}$ are around 0.05, with decreasing values of AR as $\tau_{rc}$ 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 24. In general, considering the best parameter configuration by each user, both methods (MAT-SG and MAT-SGT) present high AR values, demonstrating the effectiveness of our methods in summarizing user trajectories. MAT-SGT consistently outperforms MAT-SG in ranking user trajectories across datasets.

Table 24 – 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) |
| **Pisa** | **MAT-SG** | 0.752 | 0.800 | 0.508 | 0.500 | 0.313 | 1.000 | 0.000 | 2300 | 0 + (0 users) |
| | **MAT-SGT** | 0.742 | 0.755 | 0.498 | 0.500 | 0.319 | 1.000 | 0.000 | 2150 | 150 + (0 users) |

The *Incomplete* column shows the number of parameter configurations that did not yield an $RT$, where MAT-SGT has identified more incomplete $RT$.

### 6.3.2 RMMAT Strategy

The parameters $\tau_{rv}$ and $\tau_{rc}$ 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: (A) MAT-SG and (B) MAT-SGT.

We also present the top 10 $RT$ identified in each dataset, along with the user and parameter configuration of each computed $RT$. These analyses can help identify the users who

follow high patterns, and the high RMMAT score can highlight the power of each method in covering the representativeness of trajectories concerning the user.

#### *6.3.2.1 Foursquare-NYC dataset*

Table 25 presents 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 $\tau_{rv}$ and $\tau_{rc}$ 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 $\tau_{rv}$ and $\tau_{rc}$ both set to 0.25.

Table 25 – Average of RMMAT of user trajectories in Foursquare dataset

(A) MAT-SG

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

(B) MAT-SGT

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

Table 26 shows the top 10 *RT* computed for each method. It is interesting to note that MAT-SG achieved an RMMAT score of 0.96 for the best user (895) with its best parameter configuration, while MAT-SGT achieved an RMMAT score of 0.94 for the best user (730).

Table 26 – The top 10 computed *RT* in Foursquare dataset

(A) MAT-SG

| **user** | $\tau_{rv}$ | $\tau_{rc}$ | **RMMAT** |
|---|---|---|---|
| 895 | 0.05 | 0.05 | 0.96 |
| 730 | 0.10 | 0.05 | 0.94 |
| 754 | 0.05 | 0.25 | 0.94 |
| 207 | 0.05 | 0.10 | 0.93 |
| 1006 | 0.05 | 0.05 | 0.91 |
| 647 | 0.05 | 0.25 | 0.91 |
| 438 | 0.05 | 0.10 | 0.90 |
| 533 | 0.05 | 0.10 | 0.90 |
| 885 | 0.05 | 0.05 | 0.89 |
| 440 | 0.05 | 0.05 | 0.89 |

(B) MAT-SGT

| **user** | $\tau_{rv}$ | $\tau_{rc}$ | **RMMAT** |
|---|---|---|---|
| 730 | 0.10 | 0.05 | 0.94 |
| 895 | 0.10 | 0.05 | 0.87 |
| 207 | 0.05 | 0.10 | 0.87 |
| 754 | 0.05 | 0.25 | 0.87 |
| 365 | 0.05 | 0.05 | 0.84 |
| 647 | 0.05 | 0.25 | 0.84 |
| 69 | 0.10 | 0.05 | 0.82 |
| 440 | 0.05 | 0.05 | 0.81 |
| 438 | 0.05 | 0.10 | 0.80 |
| 673 | 0.10 | 0.05 | 0.80 |

Interestingly, although some users are common in the top-10 *RT* scores for both methods, different users were highlighted, indicating the diversity in capturing representativeness. Also, the best parameter configurations for each user vary between the two methods.

Furthermore, we can see that in these top-10 *RT*, $\tau_{rv}$ was set to 0.05 or 0.1, highlighting the tendency to obtain the best *RT* with low values. On the other hand, for $\tau_{rc}$, some *RT* were identified with high values, for example, user 647 in MAT-SG or users 754 and 647 in MAT-SGT.

#### 6.3.2.2 *Gowalla Location-Based Social Network dataset*

Table 27 insert shows results for the Gowalla dataset. The highest representativeness measures were obtained with MAT-SG (0.693) and MAT-SGT (0.624), both with $\tau_{rv}$ and $\tau_{rc}$ 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 $\tau_{rv}$ and $\tau_{rc}$ both set to 0.25.

Table 27 – Average of RMMAT of user trajectories in Gowalla dataset

(A) MAT-SG

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

(B) MAT-SGT

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

The top-10 *RT* results are detailed in Table 28. MAT-SG achieves an RMMAT score of 0.97 for the best user (36712) with its best parameter configuration, while MAT-SGT achieves an RMMAT score of 0.90 for the best user (113411). Similar to the Foursquare dataset, diverse users are highlighted in the top-10 *RT* scores for each method, showcasing the ability of each method to capture different patterns of representativeness.

Table 28 – The top 10 computed *RT* in Gowalla dataset

(A) MAT-SG

| **user** | $\tau_{rv}$ | $\tau_{rc}$ | **RMMAT** |
|---|---|---|---|
| 36712 | 0.05 | 0.25 | 0.97 |
| 18623 | 0.05 | 0.05 | 0.95 |
| 124868 | 0.05 | 0.05 | 0.95 |
| 49101 | 0.05 | 0.05 | 0.93 |
| 12681 | 0.05 | 0.05 | 0.93 |
| 107206 | 0.05 | 0.05 | 0.93 |
| 119314 | 0.05 | 0.05 | 0.93 |
| 19531 | 0.05 | 0.20 | 0.93 |
| 11205 | 0.05 | 0.05 | 0.93 |
| 113411 | 0.05 | 0.05 | 0.92 |

(B) MAT-SGT

| **user** | $\tau_{rv}$ | $\tau_{rc}$ | **RMMAT** |
|---|---|---|---|
| 113411 | 0.10 | 0.05 | 0.90 |
| 36712 | 0.05 | 0.25 | 0.89 |
| 16931 | 0.05 | 0.25 | 0.89 |
| 18623 | 0.05 | 0.05 | 0.86 |
| 119314 | 0.05 | 0.05 | 0.86 |
| 11205 | 0.05 | 0.25 | 0.86 |
| 124868 | 0.05 | 0.05 | 0.85 |
| 6321 | 0.05 | 0.25 | 0.85 |
| 39547 | 0.05 | 0.15 | 0.85 |
| 5980 | 0.05 | 0.25 | 0.83 |

It is noticeable that in the top 10 instances of repeated measures (*RT*), a value of 0.05 was frequently set for the $\tau_{rv}$ parameter, indicating a preference for obtaining the best result with low values. Conversely, for $\tau_{rc}$, certain *RT* instances were identified with high values.

#### 6.3.2.3 *Brightkite dataset*

Table 29 displays 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 $\tau_{rv}$ and $\tau_{rc}$ set to 0.05. Conversely, the lowest values were obtained by MAT-SG (0.551) with $\tau_{rv}$ and $\tau_{rc}$ both set to 0.25, and by MAT-SGT (0.298) with $\tau_{rv} = 0.25$ and $\tau_{rc} = 0.15$.

Table 29 – Average of RMMAT of user trajectories in Brightkite dataset

(A) MAT-SG

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

(B) MAT-SGT

| $\tau_{rv}$ \ $\tau_{rc}$ | **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 |

The top 10 *RT* results are detailed in Table 30. MAT-SG achieved an RMMAT score of 0.99 for the best user (7528) with its best parameter configuration, while MAT-SGT achieved an RMMAT score of 0.98 for the same user. Here, we identify that five users are the same in both method computations, all with the same parameter configuration.

Table 30 – The top 10 computed *RT* in Brightkite dataset

(A) MAT-SG

| **user** | $\tau_{rv}$ | $\tau_{rc}$ | **RMMAT** |
|---|---|---|---|
| 7528 | 0.05 | 0.25 | 0.99 |
| 662 | 0.05 | 0.25 | 0.97 |
| 22820 | 0.05 | 0.25 | 0.97 |
| 49030 | 0.05 | 0.05 | 0.96 |
| 8921 | 0.05 | 0.05 | 0.96 |
| 18841 | 0.05 | 0.25 | 0.96 |
| 26004 | 0.05 | 0.15 | 0.96 |
| 29673 | 0.05 | 0.25 | 0.96 |
| 7226 | 0.05 | 0.25 | 0.95 |
| 1952 | 0.05 | 0.05 | 0.95 |

(B) MAT-SGT

| **user** | $\tau_{rv}$ | $\tau_{rc}$ | **RMMAT** |
|---|---|---|---|
| 7528 | 0.05 | 0.25 | 0.98 |
| 22820 | 0.05 | 0.25 | 0.92 |
| 9548 | 0.05 | 0.10 | 0.92 |
| 29673 | 0.05 | 0.25 | 0.90 |
| 8921 | 0.05 | 0.05 | 0.88 |
| 11756 | 0.05 | 0.20 | 0.88 |
| 20249 | 0.05 | 0.25 | 0.87 |
| 13679 | 0.05 | 0.05 | 0.87 |
| 18841 | 0.05 | 0.25 | 0.86 |
| 43 | 0.05 | 0.25 | 0.86 |

Furthermore, we can see that in these top 10 *RT*, $\tau_{rv}$ was set as 0.05, highlighting the tendency to obtain the best *RT* with low values. On the other hand, for $\tau_{rc}$, different values are achieved. Again, the diversity in users and parameter configurations in the top 10 *RT* scores highlights the distinct capturing capabilities of MAT-SG and MAT-SGT.

#### *6.3.2.4 Pisa dataset*

Results for the Pisa dataset are shown in Table 31. The highest representativeness measures were obtained with MAT-SG (0.595) and MAT-SGT (0.621), both with $\tau_{rv}$ and $\tau_{rc}$ set to 0.05. Conversely, the lowest values were recorded for both methods (0.211 for MAT-SG and 0.327 for MAT-SGT), with $\tau_{rv}$ and $\tau_{rc}$ both set to 0.25.

The top 10 *RT* results are detailed in Table 32. MAT-SG achieved an RMMAT score of 0.90 for the best user (130) with its best parameter configuration, while MAT-SGT achieved an

Table 31 – Average of RMMAT of user trajectories in Pisa dataset

(A) MAT-SG

| $\tau_{rv}$ \ $\tau_{rc}$ | **0.05** | **0.1** | **0.15** | **0.2** | **0.25** |
|---|---|---|---|---|---|
| **0.05** | **0.595** | 0.495 | 0.390 | 0.294 | 0.231 |
| **0.1** | 0.584 | 0.483 | 0.381 | 0.287 | 0.226 |
| **0.15** | 0.573 | 0.471 | 0.369 | 0.276 | 0.217 |
| **0.2** | 0.565 | 0.461 | 0.360 | 0.271 | 0.214 |
| **0.25** | 0.559 | 0.455 | 0.355 | 0.267 | 0.211 |

(B) MAT-SGT

| $\tau_{rv}$ \ $\tau_{rc}$ | **0.05** | **0.1** | **0.15** | **0.2** | **0.25** |
|---|---|---|---|---|---|
| **0.05** | **0.621** | 0.543 | 0.471 | 0.457 | 0.452 |
| **0.1** | 0.589 | 0.504 | 0.429 | 0.405 | 0.402 |
| **0.15** | 0.551 | 0.463 | 0.404 | 0.386 | 0.372 |
| **0.2** | 0.506 | 0.422 | 0.360 | 0.363 | 0.344 |
| **0.25** | 0.476 | 0.416 | 0.355 | 0.340 | 0.327 |

RMMAT score of 1.00 for the best user (195). Here, we identify that eight users are the same in both method computations, all with the same parameter configuration.

Table 32 – The top 10 computed *RT* in Pisa dataset

(A) MAT-SG

| **user** | $\tau_{rv}$ | $\tau_{rc}$ | **RMMAT** |
|---|---|---|---|
| 130 | 0.25 | 0.25 | 0.90 |
| 439 | 0.25 | 0.25 | 0.90 |
| 443 | 0.25 | 0.25 | 0.90 |
| 99 | 0.25 | 0.25 | 0.88 |
| 480 | 0.25 | 0.10 | 0.85 |
| 543 | 0.10 | 0.10 | 0.85 |
| 744 | 0.20 | 0.10 | 0.84 |
| 195 | 0.25 | 0.10 | 0.84 |
| 672 | 0.25 | 0.15 | 0.84 |
| 191 | 0.25 | 0.10 | 0.83 |

(B) MAT-SGT

| **user** | $\tau_{rv}$ | $\tau_{rc}$ | **RMMAT** |
|---|---|---|---|
| 195 | 0.25 | 0.10 | 1.00 |
| 130 | 0.25 | 0.25 | 1.00 |
| 439 | 0.25 | 0.25 | 1.00 |
| 443 | 0.25 | 0.25 | 1.00 |
| 506 | 0.25 | 0.10 | 0.99 |
| 99 | 0.25 | 0.25 | 0.98 |
| 162 | 0.25 | 0.05 | 0.96 |
| 480 | 0.25 | 0.10 | 0.93 |
| 543 | 0.10 | 0.10 | 0.92 |
| 672 | 0.25 | 0.15 | 0.92 |

In the top-10 *RT* results, we can observe that $\tau_{rv}$ was set at 0.25, indicating a preference for high values to obtain the best *RT*. However, for $\tau_{rc}$, different values were achieved. The top 10 results for *RT* demonstrate the effectiveness of both methods in capturing the representativeness of trajectories concerning users, although the identified users and parameter configurations differ between MAT-SG and MAT-SGT.

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 33. 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.

We can observe that in some cases, there is insufficient density to determine a behavioral pattern (*Incomplete* column), where MAT-SGT has identified more incomplete *RT* across some parameter configurations.

Table 33 – 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) |
| **Pisa** | **MAT-SG** | 0.595 | 0.590 | 0.383 | 0.380 | 0.264 | 0.900 | 0.000 | 2300 | 0 + (0 users) |
| | **MAT-SGT** | 0.624 | 0.615 | 0.443 | 0.420 | 0.262 | 1.000 | 0.000 | 2150 | 150 + (0 users) |

## 6.4 DISCUSSION

We conducted a comprehensive set of experiments to assess the performance of our two trajectory summarization methods, MAT-SG and MAT-SGT, across diverse datasets with varying characteristics and parameter configurations. Our evaluation focused on their dual capabilities: effectively ranking filtered trajectories using the AR Metric and ensuring the representativeness of the computed representative data for each input dataset using the RMMAT.

The AR metric results indicated high values, particularly with the optimal parameter configurations in both MAT-SG and MAT-SGT. Values ranged between 0.687 (for MAT-SG on the Pisa dataset) and 0.954 (for MAT-SGT on the Brightkite dataset). Notably, lower values of $\tau_{rc}$ yielded better AR metric results, suggesting that $RT$ excelled in ranking trajectories when computed with larger cells, capturing more input data characteristics.

Based on the AR metric results, MAT-SGT outperformed MAT-SG in ranking user trajectories across different datasets. It achieved higher AR values than MAT-SG across various parameter configurations (Tables 20 - 23). Although MAT-SGT exhibited superior performance, the marginal overall difference emphasizes the effectiveness of both methods in ranking trajectories.

Furthermore, the highest AR values achieved with the best parameter configurations indicate the superior performance of MAT-SGT in representing diverse trajectory datasets. However, both methods produced high values, demonstrating their power to rank trajectories and ensuring the robustness of both approaches. This highlights the utility of $RT$ within the context of the input dataset, providing insights into summarization quality and representative data computation.

Based on the RMMAT results, MAT-SG demonstrates better performance in certain situations. For example, on the Brightkite dataset, MAT-SG achieved an average RMMAT score of 0.875, outperforming MAT-SGT 0.738. In general, results concerning the RMMAT showcased positive outcomes, with average values between 0.595 (for MAT-SG on the Pisa dataset) and 0.875 (for MAT-SG on the Brightkite dataset). Lower values of $\tau_{rc}$ consistently led to higher RMMAT scores across all datasets, highlighting the effectiveness of larger cell sizes in capturing representativeness.

However, it is important to note that diverse parameter configurations for different users underscore the need for adaptive parameter selection based on individual user behavior. The analysis suggests that MAT-SG showcases the best values, particularly when considering a parameter configuration around the pattern in the data. This suggests that MAT-SG may be 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.

The top 10 *RT* results demonstrate the effectiveness of both methods in capturing the representativeness of trajectories concerning users, although the identified users and parameter configurations differed between MAT-SG and MAT-SGT. This emphasizes the distinct capturing capabilities of each method. Additionally, the variability in the optimal parameter configurations highlights the importance of flexibility in parameter selection.

Both MAT-SG and MAT-SGT select parameter configurations for each user using spatial (MAT-SG and MAT-SGT) and temporal (MAT-SGT) density segmentation, analyzing aspect frequency in each segment. However, due to its dual-step density segmentation, MAT-SGT exhibited more situations with insufficient density. Consequently, more information is needed to analyze its representative data. The prevalence of different configurations is crucial since users exhibit different behavioral patterns.

Trajectory data summarization demands tailored approaches based on the specific characteristics of the data and analysis objectives. MAT-SG operates on the principle of spatial density, yielding superior results in summarizing representative aspects within specific spatial areas. It proves instrumental in scenarios where understanding spatial patterns is crucial, irrespective of the temporal sequence. For example, consider a scenario where we aim to discern the regions an individual frequents and the corresponding patterns related to each region. MAT-SG would provide insights into these spatial patterns, offering valuable information about the individual's movement across various spatial areas.

In contrast, MAT-SGT is purpose-built to solve movement patterns with a perceptive emphasis on the temporal sequence. This method excels when temporal information is critical to understanding the chronology of events or movements over time. Imagine a set of daily trajectories depicting an individual's movements on different days. Here, MAT-SGT would excel in revealing the sequence of activities the individual typically follows. For example, it could reveal that the user consistently departs from home between 7:00 am and 8:30 am on weekdays,

heads to their business in the West area between 9:00 am and 12:00 pm, and then visits a restaurant in the Center area between 12:00 pm and 1:40 pm. In this nuanced example, MAT-SGT provides a detailed temporal narrative, capturing the when and how of the routine of the user.

The choice between MAT-SG and MAT-SGT hinges on the specific objectives of the analysis. If the goal is to comprehend spatial patterns independently of temporal nuances, MAT-SG is the method of choice. On the other hand, when the temporal sequence is integral to understanding the dynamics of movements or events, MAT-SGT emerges as the preferred method. The choice between these methods represents a strategic decision, allowing analysts to tailor trajectory summarization to the unique requirements of their investigation.

These experimental evaluations provide a comprehensive and nuanced tool to understand our methods and represent filtered trajectories. Both MAT-SG and MAT-SGT demonstrate high effectiveness, and their flexibility in adapting to individual group behavior patterns is particularly valuable for personalized services and targeted interventions.

Our research has effectively demonstrated that MAT-SG and MAT-SGT are highly effective in capturing the representativeness of filtered trajectories. These methods can be applied in practical scenarios such as LBS recommendation systems, urban planning, and transportation management, where understanding filtered trajectories plays a critical role in decision-making and service optimization.

The flexibility of our methods to adapt to individual group behavior patterns is particularly valuable for personalized services and targeted interventions. By utilizing these methods, analysts can gain a deeper understanding of their data, and businesses can make more informed decisions that benefit their customers and bottom line.

### 6.4.1 Limitations

Our investigation involved a systematic experimental evaluation across multiple datasets, employing a range of parameter configurations and metrics to assess the effectiveness of our methods. The experimental results indicate that both MAT-SG and MAT-SGT exhibit strong performance in ranking filtered trajectories (AR Metric) and computing representative data for input datasets (RMMAT). Despite the positive outcomes, it is important to acknowledge several limitations of our research:

- **Parameter Sensitivity:** Our methods exhibit high sensitivity to parameter configurations, particularly $\tau_{rc}$ and $\tau_{rv}$. Lower values of $\tau_{rc}$ generally yield better results, emphasizing the critical role of parameter selection in capturing representativeness. This sensitivity necessitates meticulous parameter tuning, which may pose challenges in specific scenarios and could impact the generalizability of our methods across diverse datasets and behaviors.

- **Temporal Sequence Deficient Analysis:** The similarity analysis using MUITAS, as the similarity measure for MAT, does not account for the temporal sequence between MAT

points. The lack of consideration for temporal sequence may lead to less accurate representativeness in scenarios where the temporal order of trajectory points is significant. Based on RMMAT, MAT-SG generally demonstrates better performance in most situations. However, when comparing both methods in terms of representativeness, the use of MUITAS for similarity analysis may not provide a comprehensive assessment of their effectiveness.

- **Scalability and Computational Overhead:** The adaptability of our methods to individual user behavior patterns comes with increased computational overhead. Extensive experimentation with different parameter settings can be resource-intensive, potentially limiting the scalability of our methods for large-scale applications. The requirement for substantial computational resources may hinder the practical implementation of our approaches in real-time or resource-constrained environments.

- **Dataset Specificity:** While our methods perform well across the selected datasets, the datasets used in our evaluation may not fully represent the diversity of real-world scenarios. The effectiveness of our methods in other contexts requires further exploration. The representativeness and ranking metrics might behave differently with datasets featuring varying characteristics or noise levels, necessitating caution in extending conclusions to different scenarios.

- **Lack of Baseline Comparison:** A comparative analysis with existing MAT summarization methods is crucial to identify specific limitations of both MAT-SG and MAT-SGT. However, the absence of compatible baselines in related work hinders this comparative analysis. Recognizing areas where our methods may fall short compared to established techniques can offer valuable insights for refinement and future development.

In conclusion, while our methods show promise in trajectory summarization and representativeness measurement, these limitations underscore the need for further research and development. Addressing these challenges will be essential to enhance the robustness, scalability, and generalizability of our approaches.

# 7 CONCLUSION

This thesis has introduced a *framework* composed by two innovative methods for summarizing trajectories with multiple aspects, MAT-SG and MAT-SGT, designed to provide representative data. The previous method, the FSM-based approach (SEEP; VAHRENHOLD, 2019), had limitations in capturing temporal sequences and dealing with different aspects in their individual type. To address these shortcomings, MAT-SG and MAT-SGT consider spatial, temporal, and semantic attributes that characterize MATs. Their contribution lies in abstracting each of these dimensions according to their singularities. Another distinctive feature is mapping input MATs and representative data through a comprehensive data model, enabling persistence, querying, and pattern identification. Additionally, MAT-SGT identifies temporal sequences within movement patterns.

Trying to answer our research question, "Can we develop new algorithms for computing representative data for a set of MATs to discover relevant information and address gaps in related work by considering all aspects in MATs regarding their individually?" we tackled the trajectory summarization problem by proposing accurate methods for computing representative MAT. We propose a framework composed of two methods, called MAT-SG and MAT-SGT, that have shown promising results.

Trajectory data summarization demands a tailored approach based on specific data characteristics and analysis objectives. MAT-SG operates on spatial density, excelling in summarizing representative aspects within specific spatial areas. This method is invaluable when understanding spatial patterns is crucial, regardless of temporal sequence relevance. In contrast, MAT-SGT focuses on temporal sequence emphasis, providing detailed insights into the chronology of events or movements over time. The choice between MAT-SG and MAT-SGT depends on the analysis objectives, with MAT-SG preferred for spatial pattern comprehension and MAT-SGT for detailed temporal narratives.

Once we proposed these methods, we achieved another research question: "How much of the *RT* captures and reflects the original MATs' essence within an input dataset?". Aiming to answer this research question, we propose a representativeness measure RMMAT that refers to a measure tailored for big trajectory data with multiple aspects, aiming to quantify how much information the *RT* covers from the input dataset and how similar this *RT* is to the entire dataset.

Our exploration involved a systematic experimental evaluation across multiple datasets, employing a range of parameter configurations and metrics to assess the effectiveness of our methods. The experimental results indicate that both MAT-SG and MAT-SGT exhibit strong performance in ranking filtered trajectories (*AR Metric*) and computing representative data for input datasets (*RMMAT*).

The flexibility of our methods is highlighted by the adaptability of parameter configurations to individual user behavior. Lower values of the $\tau_{rc}$ consistently yielded better results, emphasizing the importance of parameter configuration in capturing representativeness. User-specific insights and diverse parameter configurations underscore the need for a nuanced and

adaptive approach to trajectory summarization.

Our work contributes to the trajectory data analysis research area by providing tailored methods that cater to the nuances of spatial and temporal considerations. The methods offer a nuanced tool for analysts, allowing them to choose an approach aligned with the intricacies of their data and the goals of their analysis.

As we conclude this thesis, it is evident that trajectory summarization is not a one-size-fits-all endeavor. Instead, it requires a thoughtful consideration of the specific characteristics of the data and the analytical objectives. Our methods provide a valuable step towards addressing this challenge, offering a refined and adaptable approach to trajectory summarization that can find application across diverse domains and scenarios.

It is crucial to emphasize that using a representative MAT helps data analysts gain insights into the behaviors of trajectories with multiple aspects. This allows them to understand the patterns and representative information that characterize input MATs. While our methods have shown strong performance, there are areas for future improvement, and we propose some potential future works.

The computation of a representative MAT depends on the specific purpose and requirements of a use case. The evaluation of the summarization method is also dependent on the purpose to be analyzed. Our representativeness measure RMMAT focuses on a view of similarity and covered information. In future works, we intend to employ other views to assess the representativeness of summarized MATs, such as reduced information.

Our summarization methods, MAT-SG and MAT-SGT, have effectively extracted representative MATs from trajectories with multiple aspects. In the future, we intend to use representative trajectories to analyze their impact in various scenarios. For example, we want to use representative trajectories to measure the similarity between different groups of trajectories and identify the closest group of trajectories. Additionally, we plan to use representative trajectories as input data for certain approaches. This will help us analyze the impact of using less information to be processed, like in prediction scenarios, and has potential applications in personalized recommendations, like anomaly detection and urban planning. In terms of our methods, future work aims to refine the parameter selection process to enhance the method's performance in diverse datasets and real-world scenarios.

Moreover, we acknowledge that some aspects could have relationships between them. Therefore, as a future work, we intend to improve our methods by considering dependencies between aspects, such as *price* depending on *PoI* in our running example.

Furthermore, efforts will be directed towards reducing the complexity of our methods, currently operating at $O(n^2)$ concerning the number $n$ of input points in all MATs. This complexity is primarily due to the *computeMinSpatialThreshold* function in Algorithm 1 and Algorithm 4. We will optimize key functions and ensure scalability for larger datasets to achieve this simplification.

In conclusion, trajectory summarization is a multifaceted challenge that demands precision and adaptability. Our methods provide a step forward, offering refined and adaptable

approaches that align with the intricacies of diverse datasets and analytical goals. The representative MATs derived from these methods facilitate a deeper understanding of behavioral patterns within multiple aspect trajectories, making them valuable tools for data analysts across various domains.

## 7.1 PUBLICATIONS

During this Ph.D. research period, partial results have been published as journal articles and conference papers, as follows:

- **Conference Paper: DEXA 2022 - Database and Expert Systems Applications.**

  (MACHADO; MELLO; BOGORNY, 2022a)

  Machado, V. L., Mello, R. D. S., & Bogorny, V. (2022, July). A method for summarizing trajectories with multiple aspects. In International Conference on Database and Expert Systems Applications (pp. 433-446). Cham: Springer International Publishing.

  This paper refers to the MAT-SG method, our first contribution.

- **Conference Paper: WTDBD 2022 - Workshop de Teses e Dissertações em Banco de Dados.**

  (MACHADO; MELLO; BOGORNY, 2022b)

  Machado, V. L., dos Santos Mello, R., & Bogorny, V. (2022, September). On Generating Representative Data for Multiple Aspects Trajectory Data. In Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados (pp. 98-104). SBC.

  This workshop provided a forum to present the thesis and gain valuable insights for the subsequent stages.

- **Journal articles: Revista ComInG Ed. 2022 - Communications and Innovations Gazette**

  (LUZ; MACHADO; MELLO, 2022)

  da Luz, T. O., Machado, V. L., & dos Santos Mello, R. (2022). Visual R-MAT: uma ferramenta visual de apoio a análises sob dados representativos de trajetórias de multiplos aspectos. Revista ComInG-Communications and Innovations Gazette, 6(1), 36-45.

  This paper describes a tool developed by an undergraduate student during his initial scientific research studies under my supervision.

- **Conference Paper: GeoInfo 2023 - Brazilian Symposium on Geoinformatics.**

  (MACHADO et al., 2023a)

  Machado, V. L., Portela, T. T., de Lara Machado, A., Schreiner, G. A., & dos Santos Mello, R. (2023). A method for computing representative data for multiple aspect

trajectories based on data summarization. In Brazilian Symposium on Geoinformatics (Geoinfo).

This paper refers to the MAT-SGT method, our second contribution.

- **Conference Paper: GeoInfo 2023 - Brazilian Symposium on Geoinformatics.**

  (MACHADO et al., 2023b)

  Machado, V. L., Portela, T. T., Renso, C., & dos Santos Mello, R. (2023). Towards a representativeness measure for summarized trajectories with multiple aspects. In Brazilian Symposium on Geoinformatics (Geoinfo).

  This paper refers to the additional contribution regarding a representativeness measure (RMMAT), allowing us to measure the representative data quality regarding the input data.

- **Journal paper: Geoinformatica - An International Journal on Advances of Computer Science for Geographic Information Systems.**

  (MACHADO et al., 2024)

  Machado, V. L., dos Santos Mello, R., Bogorny, V., & Schreiner, G. A. (2024). A Survey on the Computation of Representative Trajectories. GeoInformatica. Springer. 1-26.

  This paper presents a comprehensive survey and in-depth analysis of the state-of-the-art regarding to this thesis research subject.

- **Journal paper: JIDM - Journal of Information and Data Management.**

  (MACHADO et al., a)

  Machado, V. L., Portela, T. T., de Lara Machado, A., Schreiner, G. A., & dos Santos Mello, R. Towards Data Summarization of Multi-Aspect Trajectories Based on Spatio-Temporal Segmentation. JIDM.

  This document presents a MAT-SGT extended version, which was presented at the Geoinfo conference in 2023. Currently, it is in the evaluation stage.

- **Journal paper: JIDM - Journal of Information and Data Management.**

  (MACHADO et al., b)

  Machado, V. L., Portela, T. T., Vanini, L., Renso, C., & dos Santos Mello, R. A Robust Measure for Evaluating Representativeness of Summarized Trajectories with Multiple Aspects. JIDM.

  This document presents a RMMAT extended version, which was presented at the Geoinfo conference in 2023. It was accepted in the evaluation stage, and it is currently in the publication stage.

# BIBLIOGRAPHY

AGARWAL, P. K. et al. Subtrajectory clustering: Models and algorithms. In: **Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems**. [S.l.: s.n.], 2018. p. 75–87.

AHMED, M. Data summarization: a survey. **Knowledge and Information Systems**, Springer, v. 58, n. 2, p. 249–273, 2019.

AHMED, S. A. et al. Trajectory-based surveillance analysis: A survey. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 29, n. 7, p. 1985–1997, 2019.

ALMEIDA, D. R. de et al. A survey on big data for trajectory analytics. **ISPRS Int. J. Geo-Information**, v. 9, n. 2, p. 88, 2020. Disponível em: https://doi.org/10.3390/ijgi9020088.

ALVARES, L. O. et al. A model for enriching trajectories with semantic geographical information. In: SAMET, H.; SHAHABI, C.; SCHNEIDER, M. (Ed.). **15th ACM International Symposium on Geographic Information Systems, ACM-GIS 2007, November 7-9, 2007, Seattle, Washington, USA, Proceedings**. New York, NY, USA: Association for Computing Machinery (ACM), 2007. p. 22.

AMIGO, D. et al. Review and classification of trajectory summarisation algorithms: From compression to segmentation. **International Journal of Distributed Sensor Networks**, SAGE Publications Sage UK: London, England, v. 17, n. 10, p. 27, 2021.

AYHAN, S.; SAMET, H. Diclerge: Divide-cluster-merge framework for clustering aircraft trajectories. In: **Proceedings of the 8th ACM SIGSPATIAL International Workshop on Computational Transportation Science**. [S.l.: s.n.], 2015. p. 7–14.

BERNDT, D. J.; CLIFFORD, J. Using dynamic time warping to find patterns in time series. In: **Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining**. Seattle, WA: AAAI Press, 1994. (AAAIWS'94), p. 359–370.

BIAN, J. et al. A survey on trajectory clustering analysis. **CoRR**, abs/1802.06971, 2018. Disponível em: http://arxiv.org/abs/1802.06971.

BIAN, J. et al. Trajectory data classification: A review. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM New York, NY, USA, v. 10, n. 4, p. 1–34, 2019.

BLELLOCH, G. E. Introduction to data compression*. **Computer Science Department, Carnegie Mellon University**, p. 55, 2013.

BOGORNY, V.; HEUSER, C. A.; ALVARES, L. O. A conceptual data model for trajectory data mining. In: SPRINGER. **Proceedings of the 6th International Conference on Geographic Information Science**. Berlin, Heidelberg: Springer-Verlag, 2010. (GIScience'10), p. 1–15. ISBN 3642152996.

BORKOWSKI, P. The ship movement trajectory prediction algorithm using navigational data fusion. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 17, n. 6, p. 1432, 2017.

BOUDIN, F.; HUET, S.; TORRES-MORENO, J.-M. A graph-based approach to cross-language multi-document summarization. **Polibits**, Instituto Politécnico Nacional, Centro de Innovación y Desarrollo . . . , n. 43, p. 113–118, 2011.

BUCHIN, K. et al. Median trajectories. **Algorithmica**, Springer, v. 66, n. 3, p. 595–614, 2013.

BUCHIN, M.; KILGUS, B.; KöLZSCH, A. Group diagrams for representing trajectories. In: **Proceedings of the 11th ACM SIGSPATIAL International Workshop on Computational Transportation Science**. New York, NY, USA: Association for Computing Machinery, 2018. (IWCTS'18), p. 1–10. ISBN 9781450360371.

BUCHIN, M.; KILGUS, B.; KöLZSCH, A. Group diagrams for representing trajectories. **International Journal of Geographical Information Science**, Taylor & Francis, v. 34, n. 12, p. 2401–2433, 2019.

CESARIO, E.; COMITO, C.; TALIA, D. Trajectory data analysis over a cloud-based framework for smart city analytics. In: **Internet of Things Based on Smart Objects**. [S.l.]: Springer, 2014. p. 143–162.

CHANDOLA, V.; KUMAR, V. Summarization–compressing data into an informative representation. **Knowledge and Information Systems**, Springer, v. 12, p. 355–378, 2007.

CHO, E.; MYERS, S. A.; LESKOVEC, J. Friendship and mobility: User movement in location-based social networks. **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 1082–1090, 2011.

CORMEN, T. H. et al. **Introduction to algorithms**. [S.l.]: MIT press, 2009.

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

DESU, M. M. A selection problem. **The Annals of Mathematical Statistics**, JSTOR, v. 41, n. 5, p. 1596–1603, 1970.

EITER, T.; MANNILA, H. **Computing Discrete Frechet Distance**. TU Vienna - Austria, 1994.

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

ESTEBAN, J. et al. A review of data fusion models and architectures: towards engineering guidelines. **Neural Computing & Applications**, Springer, v. 14, n. 4, p. 273–281, 2005.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining**. Portland, Oregon: AAAI Press, 1996. (KDD'96), p. 226—-231.

ETIENNE, L. et al. Trajectory box plot: A new pattern to summarize movements. **Int. J. Geogr. Inf. Sci.**, Taylor & Francis, USA, v. 30, n. 5, p. 835–853, maio 2016. ISSN 1365-8816.

FENG, Z.; ZHU, Y. A survey on trajectory data mining: Techniques and applications. **IEEE Access**, IEEE, v. 4, p. 2056–2067, 2016.

FIORE, M. et al. Privacy in trajectory micro-data publishing: A survey. **Transactions on Data Privacy**, University of Skovde, v. 13, n. 2, p. 91–149, ago. 2020. ISSN 1888-5063.

FRENTZOS, E. et al. Algorithms for nearest neighbor search on moving object trajectories. **Geoinformatica**, Springer, v. 11, p. 159–193, 2007.

FURTADO, A. S. et al. Unveiling movement uncertainty for robust trajectory similarity analysis. **Int. J. Geogr. Inf. Sci.**, v. 32, n. 1, p. 140–168, 2018.

FURTADO, A. S. et al. Multidimensional similarity measuring for semantic trajectories. **Trans. GIS**, v. 20, n. 2, p. 280–298, 2016.

GAO, C. et al. Semantic trajectory representation and retrieval via hierarchical embedding. **Information Sciences**, Elsevier, v. 538, p. 176–192, 2020.

GAO, C. et al. Semantic trajectory compression via multi-resolution synchronization-based clustering. **Knowledge-Based Systems**, Elsevier, v. 174, p. 177–193, 2019.

GEORGIOU, H. et al. Moving objects analytics: Survey on future location & trajectory prediction methods. **arXiv**, abs/1807.04639, 2018.

GHODRATNAMA, S. et al. Extractive document summarization based on dynamic feature space mapping. **IEEE Access**, IEEE, v. 8, p. 139084–139095, 2020.

HESABI, Z. R. et al. Data summarization techniques for big data—a survey. In: KHAN, S. U.; ZOMAYA, A. Y. (Ed.). **Handbook on Data Centers**. New York, United States: Springer, 2015. p. 1109–1152.

HEU, J.-U.; QASIM, I.; LEE, D.-H. Fodosu: multi-document summarization exploiting semantic analysis based on social folksonomy. **Information processing & management**, Elsevier, v. 51, n. 1, p. 212–225, 2015.

HUANG, Y. et al. A survey on trajectory-prediction methods for autonomous driving. **IEEE Transactions on Intelligent Vehicles**, IEEE, v. 7, n. 3, p. 652–674, 2022.

KONG, X. et al. Big trajectory data: A survey of applications and services. **IEEE Access**, v. 6, p. 58295–58306, 2018.

LEE, J.-G.; HAN, J.; WHANG, K.-Y. Trajectory clustering: A partition-and-group framework. In: **Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: Association for Computing Machinery (ACM), 2007. (SIGMOD '07), p. 593–604. ISBN 9781595936868.

LEHMANN, A. L.; ALVARES, L. O.; BOGORNY, V. SMSM: a similarity measure for trajectory stops and moves. **Int. J. Geogr. Inf. Sci.**, v. 33, n. 9, p. 1847–1872, 2019.

LI, H. Typical trajectory extraction method for ships based on ais data and trajectory clustering. In: **2021 2nd International Conference on Artificial Intelligence and Information Systems**. [S.l.: s.n.], 2021. p. 1–8.

LIAN, J.; ZHANG, L. One-month beijing taxi gps trajectory dataset with taxi ids and vehicle status. In: **Proceedings of the First Workshop on Data Acquisition To Analysis**. [S.l.: s.n.], 2018. p. 3–4.

LUZ, T. O. da; MACHADO, V. L.; MELLO, R. dos S. Visual r-mat: uma ferramenta visual de apoio a análises sob dados representativos de trajetórias de multiplos aspectos. **Revista ComInG-Communications and Innovations Gazette**, v. 6, n. 1, p. 36–45, 2022.

MA, K. et al. What is this article about? generative summarization with the bert model in the geosciences domain. **Earth Science Informatics**, Springer, p. 1–16, 2022.

MACHADO, V. L.; MELLO, R. d. S.; BOGORNY, V. A method for summarizing trajectories with multiple aspects. In: **Int. Conf. on Database and Expert Systems Applications**. [S.l.: s.n.], 2022. p. 433–446.

MACHADO, V. L.; MELLO, R. dos S.; BOGORNY, V. On generating representative data for multiple aspects trajectory data. In: SBC. **Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados**. [S.l.], 2022. p. 98–104.

MACHADO, V. L. et al. A survey on the computation of representative trajectories. **GeoInformatica**, Springer, p. 1–26, 2024.

MACHADO, V. L. et al. A method for computing representative data for multiple aspect trajectories based on data summarization. In: **XXIV Brazilian Symposium on Geoinformatics**. [S.l.: s.n.], 2023.

MACHADO, V. L. et al. Towards a representativeness measure for summarized trajectories with multiple aspects. In: **XXIV Brazilian Symposium on Geoinformatics**. [S.l.: s.n.], 2023.

MACHADO, V. L. et al. Towards data summarization of multi-aspect trajectories based on spatio-temporal segmentation. **Journal of Information and Data Management**. Invited to extend a paper of Geoinfo 2023, in evaluation.

MACHADO, V. L. et al. A robust measure for evaluating representativeness of summarized trajectories with multiple aspects. **Journal of Information and Data Management**. Invited to extend a paper of Geoinfo 2023, in evaluation.

MARKOVIĆ, N. et al. Applications of trajectory data from the perspective of a road transportation agency: Literature review and maryland case study. **IEEE Transactions on Intelligent Transportation Systems**, IEEE, v. 20, n. 5, p. 1858–1869, 2018.

MARTINEZ, D.; CRISTOBAL, S.; BELKOURA, S. Smart data fusion: Probabilistic record linkage adapted to merge two trajectories from different sources. **Proceedings of the SESAR Innovation Days],(Dec 2018)**, 2018.

MCCLUSKEY, A.; LALKHEN, A. G. Statistics ii: Central tendency and spread of data. **Continuing Education in Anaesthesia, Critical Care and Pain**, Elsevier, v. 7, n. 4, p. 127–130, 2007.

MELLO, R. dos S. et al. MASTER: A multiple aspect view on trajectories. **Trans. GIS**, v. 23, n. 4, p. 805–822, 2019.

MOHSIN, M. et al. Improved text summarization of news articles using ga-hc and pso-hc. **Applied Sciences**, MDPI, v. 11, n. 22, p. 10511, 2021.

MUZAMMAL, M. et al. Trajectory mining using uncertain sensor data. **IEEE Access**, IEEE, v. 6, p. 4895–4903, 2017.

NARA, A. Introduction: Human dynamics research with social media and geospatial data analytics. In: ____. **Empowering Human Dynamics Research with Social Media and Geospatial Data Analytics**. Cham: Springer International Publishing, 2021. p. 1–11. ISBN 978-3-030-83010-6. Disponível em: https://doi.org/10.1007/978-3-030-83010-6_1.

PANAGIOTAKIS, C.; PELEKIS, N.; KOPANAKIS, I. Trajectory voting and classification based on spatiotemporal similarity in moving object databases. In: SPRINGER. **International Symposium on Intelligent Data Analysis**. [S.l.], 2009. p. 131–142.

PANAGIOTAKIS, C. et al. Segmentation and sampling of moving object trajectories based on representativeness. **IEEE Transactions on Knowledge and Data Engineering**, v. 24, n. 7, p. 1328–1343, 2012.

PARENT, C. et al. Semantic trajectories modeling and analysis. **ACM Comput. Surv.**, v. 45, n. 4, p. 42:1–42:32, 2013.

PETITJEAN, F.; KETTERLIN, A.; GANÇARSKI, P. A global averaging method for dynamic time warping, with applications to clustering. **Pattern Recognition**, Elsevier, v. 44, n. 3, p. 678–693, 2011.

PETRY, L. M. et al. Towards semantic-aware multiple-aspect trajectory similarity measuring. **Trans. GIS**, v. 23, n. 5, p. 960–975, 2019.

PORTELA, T. T.; CARVALHO, J. T.; BOGORNY, V. Hipermovelets: high-performance movelet extraction for trajectory classification. **International Journal of Geographical Information Science**, Taylor & Francis, v. 36, n. 5, p. 1012–1036, 2022.

PUGLIESE, C. et al. Summarizing trajectories using semantically enriched geographical context. In: **Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems**. [S.l.: s.n.], 2023. p. 1–10.

RENSO, C.; SPACCAPIETRA, S.; ZIMÁNYI, E. **Mobility Data: Modeling, Management, and Understanding**. Cambridge: Cambridge University Press, 2013.

RICHLY, K. A survey on trajectory data management for hybrid transactional and analytical workloads. In: **2018 IEEE International Conference on Big Data (Big Data)**. Seattle, WA, USA: IEEE, 2018. p. 562–569.

RODRIGUEZ, D. F.; ORTIZ, A. E. Detecting representative trajectories in moving objects databases from clusters. In: SPRINGER. **International Conference on Information Technology & Systems**. [S.l.], 2020. p. 141–151.

SEEP, J.; VAHRENHOLD, J. Inferring semantically enriched representative trajectories. In: **Proceedings of the 1st ACM SIGSPATIAL International Workshop on Computing with Multifaceted Movement Data**. New York, United States: Association for Computing Machinery, 2019. (MOVE'19), p. 1–4. ISBN 9781450369510.

SEEP, J.; VAHRENHOLD, J. K-means for semantically enriched trajectories. In: **Proceedings of the 1st ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility**. [S.l.: s.n.], 2021. p. 38–47.

SREELAKSHMI, P.; MANMADHAN, S. Image summarization using unsupervised learning. In: IEEE. **2021 7th international conference on advanced computing and communication systems (ICACCS)**. [S.l.], 2021. v. 1, p. 100–103.

SU, H. et al. A survey of trajectory distance measures and performance evaluation. **The VLDB Journal**, Springer, v. 29, n. 1, p. 3–32, 2020.

VLACHOS, M.; KOLLIOS, G.; GUNOPULOS, D. Discovering similar multidimensional trajectories. In: **Proceedings 18th international conference on data engineering**. San Jose, CA, USA: IEEE, 2002. p. 673–684.

WANG, H. et al. An effectiveness study on trajectory similarity measures. **Proceedings of the Twenty-Fourth Australasian Database Conference**, v. 137, p. 13–22, 2013.

WANG, S. et al. A survey on trajectory data management, analytics, and learning. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 54, n. 2, mar. 2021. ISSN 0360-0300.

WOOD, G. B. et al. Centroid sampling: A variant of importance sampling for estimating the volume of sample trees of radiata pine. **Forest Ecology and Management**, Elsevier, v. 36, n. 2-4, p. 233–243, 1990.

XIE, P. et al. Urban flow prediction from spatiotemporal data using machine learning: A survey. **Information Fusion**, Elsevier, v. 59, p. 1–12, 2020.

YANG, J.; WANG, P.; ZHANG, J. Benign strategy for recommended location service based on trajectory data. In: SPRINGER. **Data Science: 5th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2019, Guilin, China, September 20–23, 2019, Proceedings, Part I 5**. [S.l.], 2019. p. 3–19.

YIN, H.; WEN, Y.; LI, J. A survey of vehicle trajectory prediction based on deep-learning. In: IEEE. **2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE)**. [S.l.], 2023. p. 140–144.

YING, X.; XU, Z.; YIN, W. G. Cluster-based congestion outlier detection method on trajectory data. In: IEEE. **2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery**. [S.l.], 2009. v. 5, p. 243–247.

ZHENG, Y. Trajectory data mining: an overview. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM New York, NY, USA, v. 6, n. 3, p. 1–41, 2015.

ZHOU, F. et al. Trajectory-user linking via variational autoencoder. In: **IJCAI**. [S.l.: s.n.], 2018. p. 3212–3218.