On Generating Representative Data for Multiple Aspects Trajectory Data

Vanessa Lago Machado[1,2], **Ronaldo dos Santos Mello**[1], **Vânia Bogorny**[1]

[1]PPGCC-INE - Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brasil

[2]Instituto Federal Sul-Rio-Grandense (IFSul)
Passo Fundo – RS – Brasil

vanessa.machado@gmail.com,{r.mello,vania.bogorny}@ufsc.br

Resumo. *Tarefas de mineração e analise de dados de trajetórias têm sido amplamente estudadas nos últimos anos. Essas tarefas são complexas devido ao grande volume de dados gerados e sua heterogeneidade. Uma solução para minimizar esses problemas é a sumarização desses dados, visando gerar dados representativos. Poucos trabalhos na literatura se direcionam a estas soluções, e não foi encontrada nenhuma que considere todas as dimensões de uma trajetória (espacial, temporal e ilimitados aspectos semânticos), analisando as peculiaridades e singularidades de cada aspecto. Essa tese de doutorado propõe um método baseado em uma grade espacial para sumarização de trajetórias de múltiplos aspectos, chamado MAT-SG. Suas principais contribuições são: (i) segmentação das trajetórias em uma grade espacial de acordo com a dispersão dos pontos; (ii) a partir de um conjunto de trajetórias de entrada, uma trajetória representativa é gerada como uma sequência de pontos representativos com valores representativos para cada dimensão, considerando as particularidades de cada tipo de aspecto. Um exemplo demonstra o potencial da proposta, sendo avaliadas a redução de volume e a acurácia da sumarização.*

1. General information

- **Level:** Doctorate degree
- **Student:** Vanessa Lago Machado
- **Advisor:** Ronaldo dos Santos Mello
- **Co-Advisor:** Vânia Bogorny
- **Admission:** August 2019
- **Qualification Exam:** March 2022
- **Defense Forecast:** August 2023
- **Completed Steps:** All mandatory course credits (2019 - 2020), Progress seminar (2021), Qualification Exam (2022), Bibliographic Review, Problem Statement, Proposal of a summarization method based on Spatial Grid (MAT-SG).
- **Future Steps:** (i) MAT-SG: evaluation and proposal of a representativeness measure for a representative trajectory; (ii) proposal of a method for MAT summarization based on embeddings (focus on semantic dimension) and validation; (iii) thesis writing and defense; (iv) publications of the results.
- **Publication:** A Method for Summarizing Trajectories with Multiple Aspects, In: Int. Conf. on Database and Expert Systems Applications (DEXA 2022).

2. Context and Motivation

Object movement is collected today by different sensors and devices, being represented as a sequence of points with a spatial position at a time instant, which is called *raw trajectory* [Erwig et al. 1999]. More recently, we have new trajectory definitions: when the movement of the object is enriched with single semantic information, we have a *semantic trajectory* [Parent et al. 2013], and a *multiple aspect trajectory (MAT)* [Mello et al. 2019] when the enrichment considers many semantic contexts.

Due to the big data volume and variety collected, trajectory data management has become more complicated. In addition, many approaches consider this vast volume of data to analyze patterns and carry out their solutions, such as prediction, clustering, and classification algorithms. However, this big data management of trajectory data still is a challenge in the literature [Almeida et al. 2020, Wang et al. 2021, Xie et al. 2020].

Understanding the pattern of a trajectory dataset can help data analysts make better decisions. Recommendation systems, for example, deal with the analysis of users' behaviors to help them find products of interest or suggest actions that will let he/she more healthy/satisfied. As an alternative to the management challenge, a trajectory dataset can be summarized into representative data, reducing the data volume to be managed and analyzed. Then, *trajectory data summarization* is helpful to reduce the complexity of the data to be processed.

Some surveys point out that semantic trajectory data summarization is an open issue [Fiore et al. 2020, Wang et al. 2021]. This lack of works is probably due to the complexity of these data, as different semantic contexts may coexist and be related to parts of a trajectory, making data summarization tasks more challenging. Given this motivation and the literature lack about the summarization of MATs, this PhD Thesis intends to deal with this problem by answering the following research question:

How to summarize trajectories data enriched with many semantic contexts (MATs) in order to generate a trajectory with representative information?

Our hypothesis to justify this research question as a relevant research topic is that MAT data summarization allows a generalization of the data, which reduces the volume of data to be managed and/or analyzed and even stored when we think about materialized views of data. Additionally, representative MAT information through data summarization facilitates the visualization of these data and, as a consequence, their investigation.

This PhD Thesis proposal aims to deal with this open issue considering that the concept of MAT, as well as their data management, is a brand new research topic. The main objective of this thesis is to propose *a pioneer approach for MAT summarization that generates a representative MAT from a set of MATs*. In short, we intend to develop a customized computational process to generate a representative MAT based on the selection of the relevant aspects from an input MAT dataset an expert user intends to summarize.

Several strategies can be proposed to generate this representative MAT. Our first contribution at this moment is the *MAT-SG (Multiple Aspect Trajectory Summarization based on a spatial Grid)* method. It is based on a spatial grid that covers the set of input MATs. We generate a representative point for all points in the same grid cell. In turn, a *representative MAT* is generated from the sequence of representative points. Our method reduces the volume of MATs data with low accuracy loss.

The rest of this paper is organized as follows. Next section discusses related work. Section 4 details and evaluates MAT-SG. Section 5 concludes the paper.

3. Related Work

Different approaches summarize trajectories generating three types of representative data: *(i)* an entire trajectory [Buchin et al. 2013, Seep and Vahrenhold 2019, Li 2021]; *(ii)* subtrajectory [Lee et al. 2007, Panagiotakis et al. 2012, Agarwal et al. 2018, Buchin et al. 2019, Rodriguez and Ortiz 2020]; and *(iii)* region (representative area) [Gao et al. 2019]. This representative data can be output as a single object or a set of objects.

Some studies consider only the spatial dimension to generate the representative data [Lee et al. 2007, Buchin et al. 2013]. Instead, most of them also analyze the temporal dimension [Panagiotakis et al. 2012, Agarwal et al. 2018, Gao et al. 2019, Buchin et al. 2019]. Only two studies consider some aspects of the semantic dimension while summarizing trajectories [Seep and Vahrenhold 2019, Li 2021]. The second one refers to the vessel scenario, where semantic attributes are defined a priori, given by the vessel speed and direction besides location and position time. The first one considers a spatial trajectory annotated with additional information, and all attributes of the points are treated as *spatial* or *non-spatial* value.

Then, only one method considers the semantic dimension as multiple aspect data, i.e., has no limitation in the number of semantic aspects [Seep and Vahrenhold 2019]. However, semantic data are not analyzed individually as categorical or numeric data to understand the patterns and influence of each aspect in the generation of the representative data. Additionally, a strong limitation of this work is the lack of method details, as it is a short paper.

An open issue identified in the literature is the lack of a summarization method that generates a representative MAT for a set of MATs, considering unlimited semantic information treated as categorical or numerical data to understand the patterns and their influence on each data in the representative trajectory. In addition, only one study maintains mapping information to the input data, but this is not detailed [Gao et al. 2019].

4. Proposal

Our first proposal is a novel method for summarizing MAT data: *MAT-SG (Multiple Aspect Trajectory Summarization based on a spatial Grid)* [Machado et al. 2022]. MAT-SG is inspired by the literature gap regarding MAT summarization. We assume the input MATs were already filtered by some criteria, so the representative MAT denotes the main behavior of these input MATs considering the spatial density and frequency of each aspect attribute value.

Fig. 1 gives an overview of MAT-SG. The trajectory data previously filtered is given as input. Then, we standardize the input data representation using a data model. Next, the method holds two internal steps: *(i) spatial segmentation*; and *(ii) data summarization*. In order to identify patterns by spatial density, we chose to segment MAT points into a grid in the first step. Clusters of nearby points in the same cell are then generated.

The second step, in turn, generates the representative MAT (rt). It computes a representative point (p_r) for each relevant cell summarizing each dimension. Then, the rt is given as output.

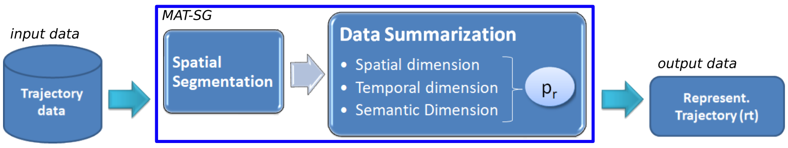


Figure 1. MAT-SG overview.

Our proposed data model maintains rt points as well as their mappings to the input MAT points since MAT-SG holds mapping data between the input MATs and the representative MAT. A p_r is generated from many MAT points, and a relationship between p_r and the corresponding MAT points is modeled. For sake of paper space, we do not detail the proposed data model[1].

The first MAT-SG step segments the points of the input MATs over a grid of squared cells. We dimension the cell size according to the dispersion of the MATs points. Next, we identify relevant cells, i.e., the cells with a predefined minimum number of points. These cells are then considered relevant to hold a p_r, otherwise they are assumed as a weak representative point.

The second MAT-SG step receives MATs segmented on a spatial grid as input and summarizes points in the same cell to generate a p_r. The p_r generation takes into account the analysis of the three MAT dimensions for all points in the cell. The summarization of each dimension is added to p_r. For spatial dimension, the centroid point is computed. Regarding temporal summarization, we discover and rank the most significant time intervals in a cell. Regarding semantic summarization, we divide it into two types: *(i)* categorical (*e.g.*, mean of transportation and weather condition) and *(ii)* numerical (*e.g.*, temperature and air humidity). For categorical types, we rank the semantic values that best represent the behavior of the cell points, and for numerical values, we compute the *median* value.

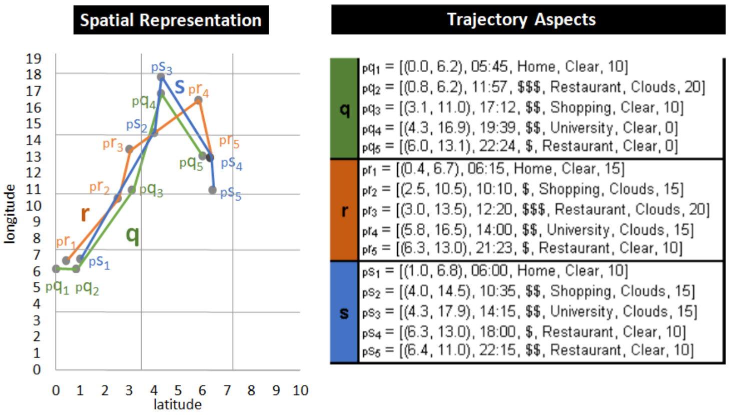
Different from related work, MAT-SG is a detailed data summarization method that generates an rt from a set of MATs, i.e., a trajectory enriched with unlimited semantic information represented as categorical or numerical data. Our approach treats semantic data individually, allowing us to understand the patterns and influence of each data in rt. It also holds mapping data between the input MATs and rt.

4.1. Running Example

To exemplify the application of MAT-SG, we present a running example. Let $T = \langle q, r, s \rangle$, where $q = \langle p_{q1}, p_{q2}, ..., p_{q_n} \rangle$, $r = \langle p_{r_1}, p_{r_2}, ..., p_{r_m} \rangle$ and $s = \langle p_{s_1}, p_{s_2}, ..., p_{s_t} \rangle$, the input MATs of some individuals. Fig. 2 presents them and some related aspects: *price* they spend in a PoI, the *PoI* itself, *weather* condition and *rain precipitation*.

Let $|T.points| = 15$ and a relevant cell must contain at least 10% of points (2 points). Then, data summarization occurs at cells containing more than one point. Fig. 3 (a) shows T segmentation into a grid of cells, and Fig. 3 (b) shows the resulting $rt =$

[1] The conceptual model for MAT-SG is detailed in [Machado et al. 2022]



Trajectory Aspects

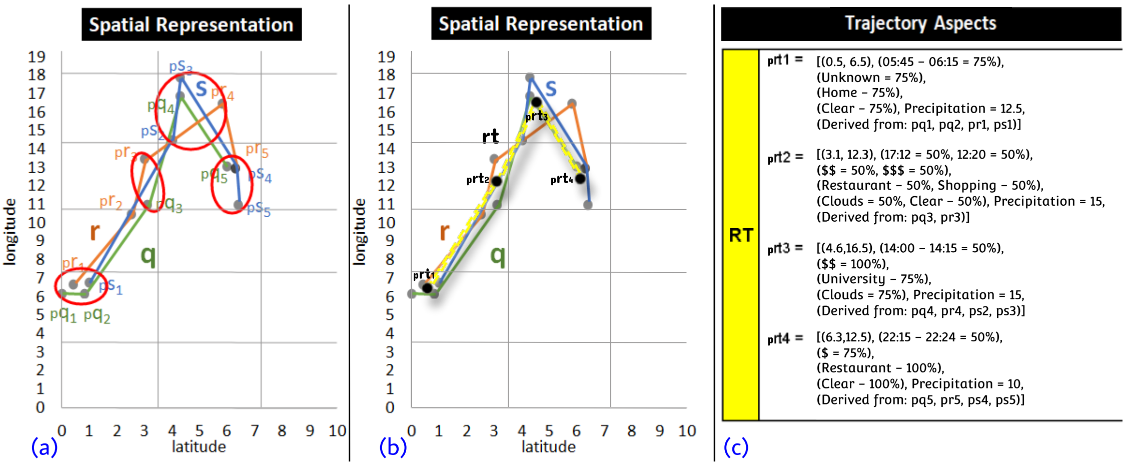
q	pq₁ = [(0.0, 6.2), 05:45, Home, Clear, 10]
	pq₂ = [(0.8, 6.2), 11:57, $$$, Restaurant, Clouds, 20]
	pq₃ = [(3.1, 11.0), 17:12, $$, Shopping, Clear, 10]
	pq₄ = [(4.3, 16.9), 19:39, $$, University, Clear, 0]
	pq₅ = [(6.0, 13.1), 22:24, $, Restaurant, Clear, 0]

r	pr₁ = [(0.4, 6.7), 06:15, Home, Clear, 15]
	pr₂ = [(2.5, 10.5), 10:10, $, Shopping, Clouds, 15]
	pr₃ = [(3.0, 13.5), 12:20, $$$, Restaurant, Clouds, 20]
	pr₄ = [(5.8, 16.5), 14:00, $$, University, Clouds, 15]
	pr₅ = [(6.3, 13.0), 21:23, $, Restaurant, Clear, 10]

s	ps₁ = [(1.0, 6.8), 06:00, Home, Clear, 10]
	ps₂ = [(4.0, 14.5), 10:35, $$, Shopping, Clouds, 15]
	ps₃ = [(4.3, 17.9), 14:15, $$, University, Clouds, 15]
	ps₄ = [(6.3, 13.0), 18:00, $, Restaurant, Clear, 10]
	ps₅ = [(6.4, 11.0), 22:15, $$, Restaurant, Clear, 10]

Figure 2. Sample data with point aspects information for trajectories q, r and s.

$\langle p_{rt_1}, p_{rt_2}, ..., p_{rt_k} \rangle$, where the yellow line shows the spatial dimension summarization. The detailed output is illustrated in Fig. 3 (c).



Trajectory Aspects

RT	prt1 =	[(0.5, 6.5), (05:45 – 06:15 = 75%), (Unknown = 75%), (Home – 75%), (Clear – 75%), Precipitation = 12.5, (Derived from: pq1, pq2, pr1, ps1)]
	prt2 =	[(3.1, 12.3), (17:12 = 50%, 12:20 = 50%), ($$ = 50%, $$$ = 50%), (Restaurant – 50%, Shopping – 50%), (Clouds = 50%, Clear – 50%), Precipitation = 15, (Derived from: pq3, pr3)]
	prt3 =	[(4.6,16.5), (14:00 – 14:15 = 50%), ($$ = 100%), (University – 75%), (Clouds = 75%), Precipitation = 15, (Derived from: pq4, pr4, ps2, ps3)]
	prt4 =	[(6.3,12.5), (22:15 – 22:24 = 50%), ($ = 75%), (Restaurant – 100%), (Clear – 100%), Precipitation = 10, (Derived from: pq5, pr5, ps4, ps5)]

Figure 3. A spatial segmentation (a) and the resulting rt (b, c)).

For the temporal dimension, we find out some significant intervals. This is the case of p_{rt_1}, where only one interval is identified and represents 75% of the cell points. For p_{rt_2}, for example, we have two punctual occurrences, since the time difference between p_{q_3} and p_{r_3} is very high, and these occurrences do not generate a significant interval.

For the semantic dimension, we compute the median of *rain precipitation*, which is numeric data. For the categorical data, we define a frequency ranking considering representativeness values (at least 30% in this case). This is the case for *price, POI,* and *weather*. These rankings can be useful to the analyst. For p_{rt_1}, for example, we see that, in most of the cases, the location refers to *Home*, so we can strongly suppose that this is a residential area.

4.2. Experimental Evaluation

In a preliminary experiment, we evaluate MAT-SG in terms of *data reduction* and *accuracy* compared to the input data. We use this information to analyze and define the best cell size to segment input MATs and generate rt. We analyze $|rt|$ for evaluating rate reduction, i.e., the number of rt points. In terms of accuracy, we analyze how much rt represents the input MATs by evaluating two criteria: *(i)* covered MAT points (T^c); and *(ii)* information on the covered MATs (R_M).

The literature mentions the lack of a well-defined measure for quantitative evaluation of how an rt represents all the input data [Panagiotakis et al. 2012, Machado et al. 2022]. So, we propose a *proximity metric* for evaluating the accuracy that informs how much each input trajectory is close to rt in all dimensions. We base this metric on the *match* function for MATs called *MUITAS* [Petry et al. 2019]. MUITAS is the state-of-the-art w.r.t. MAT similarity measure that measures the similarity between two MATs quantifying the distance between their points. Based on MUITAS, for computing the proximity of each input trajectory to rt, we define a proximity score. Finally, we compute our representativeness measure (R_M) metric as a median of proximity scores. For sake of paper space, we do not detail R_M.

Our method is implemented in Java, and all experiments ran on a Dell Inspiron laptop with an Intel Core i5 processor and 16 GB memory. For our running example (detailed in Section 4.1), rt covers 93% of all input MAT points (T^c = 14). W.r.t. covered information, rt captures 65% of all input MATs considering all aspects. In terms of volume, rt represents a reduction of $73,33\%$ of all input MATs ($|rt| = 4$ and $|T.points| = 15$). As the average size of all input MAT is 5, $|rt|$ is close to the size of each one.

5. Contributions and Future Works

This doctoral work intends to contribute to the open issue regarding summarizing MATs. We introduce a pioneering method to summarize MATs named MAT-SG[2]. MAT-SG considers spatial, temporal, and different semantic attributes that characterize MATs, abstracting each one of these dimensions according to their singularities. Another differential is the mapping between input MATs and the representative MAT through a data model. It allows persistence and querying of representative MATs, as well as their origins, and it also allows the analyst to identify patterns in the data and the representativeness of some MAT points. Preliminary experiments demonstrated that MAT-SG is a promising method. It was not possible to compare MAT-SG against the unique close baseline [Seep and Vahrenhold 2019] as its source code was not available, and its presented evaluation did not show the output data to allow us a comparison with our output.

This Thesis is under development, and the next steps are: *(i)* to improve MAT-SG by considering dependencies between aspects; *(ii)* to evaluate MAT-SG considering the dimensions summarized in other works, in order to validate the representativeness measure; *(iii)* to propose and develop more flexible summarization methods for MATs, since MAT-SG gives priority to the spatial dimension; *(iv)* to submit papers to conferences and journals; *(v)* Doctoral Thesis defense.

References

Agarwal, P. et al. (2018). Subtrajectory Clustering: Models and Algorithms. In *SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 75–87.

Almeida, D. R. et al. (2020). A Survey on Big Data for Trajectory Analytics. *ISPRS Int. J. Geo-Information*, 9(2):88.

[2] avalaible in `https://github.com/vanessalagomachado/MAT-SG/tree/master`.

Buchin, K. et al. (2013). Median Trajectories. *Algorithmica*, 66(3):595–614.

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

Erwig, M. et al. (1999). Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases. *GeoInformatica*, 3(3):269–296.

Fiore, M. et al. (2020). Privacy in Trajectory Micro-data Publishing: A Survey. *Transactions on Data Privacy*, 13:91–149.

Gao, C. et al. (2019). Semantic Trajectory Compression via Multi-resolution Synchronization-based Clustering. *Knowledge-Based Systems*, 174:177–193.

Lee, J.-G., Han, J., and Whang, K.-Y. (2007). Trajectory Clustering: A Partition-and-Group Framework. In *SIGMOD International Conference on Management of Data*, page 593–604. ACM.

Li, H. (2021). Typical Trajectory Extraction Method for Ships Based on AIS Data and Trajectory Clustering. In *2nd International Conference on Artificial Intelligence and Information Systems*, pages 1–8.

Machado, V. L., Mello, R. d. S., and Bogorny, V. (2022). A Method for Summarizing Trajectories with Multiple Aspects. In *International Conference on Database and Expert Systems Applications*, pages 433–446. Springer.

Mello, R. et al. (2019). MASTER: A Multiple Aspect View on Trajectories. *Transactions in GIS*, 23(4):805–822.

Panagiotakis, C. et al. (2012). Segmentation and Sampling of Moving Object Trajectories Based on Representativeness. *IEEE Trans. on Know. and Data Eng.*, 24(7):1328–1343.

Parent, C. et al. (2013). Semantic Trajectories Modeling and Analysis. *ACM Comput. Surv.*, 45(4):42:1–42:32.

Petry, L. M. et al. (2019). Towards Semantic-aware Multiple-aspect Trajectory Similarity Measuring. *Transactions in GIS*, 23(5):960–975.

Rodriguez, D. F. and Ortiz, A. E. (2020). Detecting Representative Trajectories in Moving Objects Databases from Clusters. In *International Conference on Information Technology & Systems*, pages 141–151. Springer.

Seep, J. and Vahrenhold, J. (2019). Inferring Semantically Enriched Representative Trajectories. In *SIGSPATIAL International Workshop on Computing with Multifaceted Movement Data*, pages 1–4. ACM.

Wang, S., Bao, Z., Culpepper, J. S., and Cong, G. (2021). A Survey on Trajectory Data Management, Analytics, and Learning. *ACM Comput. Surv.*, 54(2).

Xie, P. et al. (2020). Urban Flow Prediction from Spatiotemporal Data Using Machine Learning: A Survey. *Information Fusion*, 59:1–12.