# MAT-CA: a tool for Multiple Aspect Trajectory Clustering Analysis

Yuri Santos
Universidade Federal de Santa Catarina
yuri.nassar@posgrad.ufsc.br

Ricardo Giulliani
Universidade Federal de Santa Catarina
ricardogiuliani@outlook.com

Tarlis Portela
Instituto Federal do Paraná (IFPR)
tarlis.portela@ifpr.edu.br

Chiara Renso
ISTI - CNR
chiara.renso@isti.cnr.it

Jônata Carvalho
Universidade Federal de Santa Catarina
jonata.tyska@ufsc.br

## ABSTRACT

Multiple aspect trajectory (MAT) is a relevant concept that enables mining interesting patterns moving objects for different applications. This new way of looking at trajectories includes a semantic dimension, which presents the notion of aspects that are relevant facts of the real world that add more meaning to spatio-temporal data. The high dimensionality and heterogeneity of these data makes clustering a very challenging task both in terms of efficiency and quality. The present demo offers a tool, called MAT-CA, to support the user in the clustering task of MATs, specifically for identifying and visualizing the hidden patterns. The MAT-CA join into the same tool a multiple aspects trajectories clustering method and visual analysis of the results. We illustrate the use of the tool for offering both clustering output visualization and statistics.

## CCS CONCEPTS

• **Human-centered computing → Visualization systems and tools**; • **Computing methodologies → Cluster analysis**;

## KEYWORDS

Trajectory clustering, Clustering visualization, Clustering analysis

**ACM Reference Format:**
Yuri Santos, Ricardo Giulliani, Tarlis Portela, Chiara Renso, and Jônata Carvalho. 2023. MAT-CA: a tool for Multiple Aspect Trajectory Clustering Analysis. In *1st ACM SIGSPATIAL International Workshop on Methods for Enriched Mobility Data: Emerging issues and Ethical perspectives 2023 (EMODE'23), November 13, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3615885.3628009

## 1 INTRODUCTION AND MOTIVATION

Mobility data generation can be generated from different sources such as GPS devices, wearables, sensor networks, and social networks, which enables trajectory recording of humans, vehicles, animals, etc [5]. Mobility data representation and analysis have many applications in real-life, such as studying the movement of people [9], vehicles [23], ships [20], hurricanes [2] and animals [11]. A typical analysis task is the clustering of trajectories like examples of route planning [8], user profiling [10], the detection of outlier behaviour [13], or extracting common patterns from data[14].

The advancement of the Internet of Things has allowed the extraction of numerous data other than spatial-temporal. The additional data used to enrich raw trajectory data are often called *aspects*. The aspects contribute to enriching the semantic dimensions of mobility data which can contain any data type and format. This complex type of trajectory is called Multiple Aspect Trajectory (MAT) [12]. The enriched dimension can be associated with a moving object, the entire trajectory, or a single trajectory point that contains heterogeneous data types. For instance, a certain point of a trajectory that is found in a restaurant may contain, in addition to the aspect that defines the type of place visited, aspects such as price range, user evaluation, and opening hours, among others.

This highly complex and heterogeneous data type, despite very useful, presents additional challenges for data mining tasks. Clustering this type of data objects is a very challenging task both in terms of data volume and heterogeneity. The high number of attributes and the sequential nature of the data can increase the computational complexity and affect the cluster results consistency. This scenario makes it important that clustering methods proposed for MATs adopt multidimensional similarity metrics or other strategies that allow capturing such data heterogeneity [25]. Recently, Varlamis et al. [21] proposed a MAT similarity measure with hierarchical clustering by including a multi-vector representation of MATs that enables performing cluster analysis on them. Varlamis et al. [21] is one of the first works that dig deeper into the MAT clustering task. The most recent method for clustering MAT data is called MAT-Tree [19] and brings a frequent-based and tree-based approach for clustering MATs. MAT-Tree groups multiple aspect trajectories using the hierarchical clustering algorithm based on a decision tree structure that chooses the best aspect to branch and group the most similar trajectories with respect to the frequency in which the values of the different aspects happen in each trajectory.

It is noteworthy that, differently from supervised classification tasks, objectively analyzing clustering results is challenging and very domain-dependent.Internal evaluation metrics for clustering such as Sum of Square Error (SSE) [3], Silhouette coefficient [15], Calinski-harabasz [22] and Davies-Bouldin index [24], just to cite a few, are comparative metrics that do not provide insights about the semantics of the final clustering outcomes. Analyzing the semantics of the clustering process often involves performing exploratory

*EMODE'23, November 13, 2023, Hamburg, Germany*
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0347-8/23/11.
https://doi.org/10.1145/3615885.3628009

data analysis, i.e., using visualization techniques and descriptive statistics, and validating the patterns extracted with experts from the application domain. In this context, tools that could provide a simple and effective interface for applying novel clustering methods and facilitating the analysis of the outcomes could be useful for fostering the application of these methods to specific domains.

There are tools focusing on trajectory data visualization such as SCIKIT-MOBILITY[1] and GEOPANDAS[2]. Both provides management and analysis tools for raw trajectories.Another important tool is AutoMATize [18] which provides a simple software interface for MAT classification and visualization of the results. Besides, there are works that focus on proposing new MAT clustering techniques [19, 21], while other works focus only on spatio-temporal [7] visualization task. Few works proposed tools for dealing with both clustering and visualization tasks, but they focus on spatial [6] or spatial-temporal [1, 4, 17] dimensions. The MAT clustering analysis, due to the intrinsic complexity of the data, needs more specific tools that could offer visualization associated with patterns discovered from multidimensional data. Moreover, another fundamental issue is how to make understandable the patterns extracted from such high dimensional data. Clustering and visualization of trajectories in the same task faces the following main challenges:

1) Handle heterogeneous data;
2) Cluster trajectory data that is noisy or contains outliers;
3) Visualize the clustered data in a way that is easy to understand and interpret, and it is easy enough for application-domain experts, not experts on computer science, to use.

This demo paper approaches these challenges by introducing the tool MAT-CA (Mutiple Asptect Trajectory Clustering Analysis), an interactive jupyter notebook[3][4] tool that provides a friendly MAT clustering and visualization analysis. Our proposal uses the MAT-Tree clustering algorithm to identify clusters and then integrate it into the visualization step. More specifically, our tool provides the following features: 1) an interface for easily applying the MAT-tree clustering method[19] and 2) a friendly interface for summarizing and visualizing the results. For example, this tool can be used by an application-domain expert aiming to visualize the patterns extracted by MAT-tree from multiple aspect trajectories in order to validate them and to extract useful insights from their semantics.

## 2 INTERACTIONS AND DEMONSTRATION SCENARIO

MAT-CA[5] is a tool that uses widget functionalities to create interaction for clustering analysis using different views (interfaces). The prototype is tested and validated with the public Foursquare NY dataset [16]. Petry et al. [16] enriched the original Foursquare dataset [26] to evaluate MAT similarity measures. In summary, the original dataset proposed by Yang et al. [26] contains 227,428 check-ins of 1,083 different users. The check-ins were collected between April 2012 and February 2013, and each check-in is composed of a timestamp with its corresponding Foursquare venue ID. After to preprocessing, the final dataset (csv file with nine attributes) contains a total of 66,962 check-ins distributed in 3,079 weekly trajectories of 193 different users, with an average length of 22 points per trajectory and an average of 16 trajectories per user.

MAT-CA starts by running the clustering method and then the Python interface is launched to present the clustering result. This tool just supports the MAT-Tree [19] clustering algorithm at this moment, where the hyperparamenters are user-defined and application-dependent. We plan to extend the list of algorithms in the future. For sake of simplicity, the MAT-CA can be seen as post-processing tool which comprehend seven interfaces: i) Tree view, ii) Trajectory view, iii) Trajectory similarity, iv) Exploratory data analysis, v) Dataset exploration, vi) Sankey aspect relevance and vii) Heatmap. The user can interact with these interfaces (views) through dropdowns and inputs working as filters (node, number of rows, and user) on the jupyter notebook. We describe the filters and each view in the following sections. Besides, we emphasize that there is no specific order in the use of these views and they can be employed as the user needs. For instance, after the clustering method is applied, the application-domain expert can navigate on the tree visualization or check the heatmap information of the clustering result. We highlight that the tool works locally, thereby the scalability is limited to the local computer resources. It is a limit that the user has to manage as large datasets will be very hard to load and visualize. However, this limitation is removed when the tool is hosted on a high-performance computing server. Since jupyter notebooks are hosted and served using the HTTP protocol, it is easy to host the tool in a high-performance web server.

### 2.1 Tree

This interface shows the clustering result using the directed graph visualization. The *tree view* illustrates the most relevant aspects selected during the application of the clustering technique as well as the number of trajectories in each cluster. Figure 1 shows the clustering result containing four clusters (leaf nodes) with *Travel & Transport* and *Food* aspects as the most relevant to split the set of trajectories. We do not interact with filters in this view once it shows just the clustering tree. Besides, the analyst can check which threshold value was used to split the set of trajectories and the number of trajectories that match the path rule.

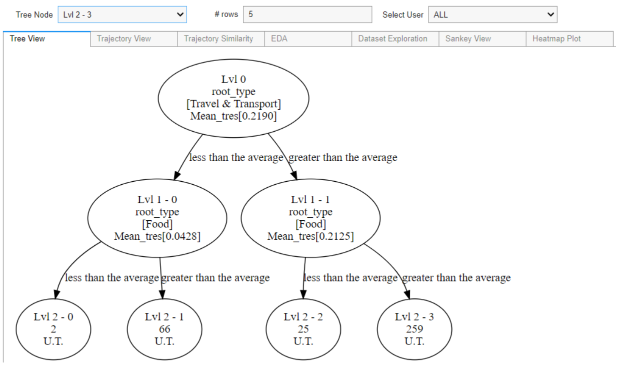
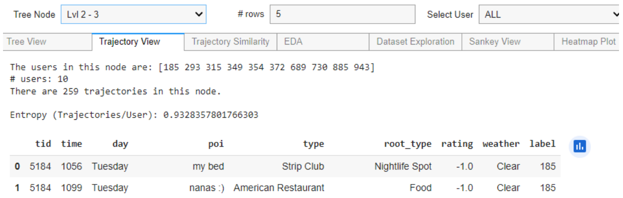

Tree Node Lvl 2 - 3
# rows 5
Select User ALL

**Figure 1: Tree view**

[1] https://github.com/scikit-mobility/scikit-mobility
[2] https://doi.org/10.5281/zenodo.705645
[3] Jupyter Notebook is a web-based interactive computing platform
[4] https://jupyter.org/
[5] Source code: https://github.com/Yuri-Nassar/mat-clus-analysis-demo

## 2.2 Trajectory

The *trajectory view* shows general information of a tree node. Figure 2 depicts the summary for node 3 at level 2 which clustered 259 trajectories distributed between 10 users plus the entropy (diversity) information. In addition, it presents the short view of the dataset related to the trajectories clustered into this node. Besides, the analyst can use the filters to check some details such as unique users, entropy, number of trajectories of a user, and the dataset information. Thus, it can support the analyst to check other aspects in the given set of trajectories grouped regarding the tree node.


Tree Node: Lvl 2 - 3 # rows: 5 Select User: ALL

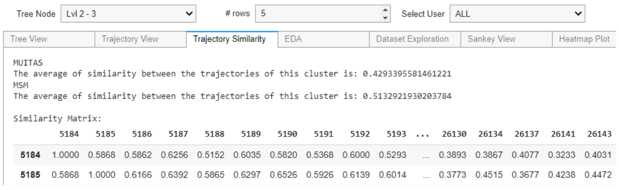
```
The users in this node are: [185 293 315 349 354 372 689 730 885 943]
# users: 10
There are 259 trajectories in this node.

Entropy (Trajectories/User): 0.9328357801766303
```

| | tid | time | day | poi | type | root_type | rating | weather | label |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 5184 | 1056 | Tuesday | my bed | Strip Club | Nightlife Spot | -1.0 | Clear | 185 |
| 1 | 5184 | 1099 | Tuesday | nanas :) | American Restaurant | Food | -1.0 | Clear | 185 |

Figure 2: Trajectory view

## 2.3 Trajectory similarity

This interface shows the average similarity scores in the node as well as the similarity dataframe. Figure 3 shows the average similarity score and the similarity matrix. We highlight that MAT-Tree does not compute the similarities for clustering, instead, it computes the aspects frequency matrix. Indeed, we use the precomputed similarity matrices to assesses the clustering quality. We have available four precomputed matrices which were generated using the MUITAS, MSM, EDR and LCSS similarity measures [16, 21]. Thus, it is another view to support the analyst comprehension about how much similar is the trajectory set for a given node.


Tree Node: Lvl 2 - 3 # rows: 5 Select User: ALL

```
MUITAS
The average of similarity between the trajectories of this cluster is: 0.4293395581461221
MSM
The average of similarity between the trajectories of this cluster is: 0.5132921930203784

Similarity Matrix:
```

| | 5184 | 5185 | 5186 | 5187 | 5188 | 5189 | 5190 | 5191 | 5192 | 5193 | ... | 26130 | 26134 | 26137 | 26141 | 26143 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5184 | 1.0000 | 0.5868 | 0.5862 | 0.6256 | 0.5152 | 0.6035 | 0.5820 | 0.5368 | 0.6000 | 0.5293 | ... | 0.3893 | 0.3867 | 0.4077 | 0.3233 | 0.4031 |
| 5185 | 0.5868 | 1.0000 | 0.6166 | 0.6392 | 0.5865 | 0.6297 | 0.6526 | 0.5926 | 0.6139 | 0.6014 | ... | 0.3773 | 0.4515 | 0.3677 | 0.4238 | 0.4472 |

Figure 3: Trajectory similarity view

## 2.4 Exploratory data analysis

We aim to show in this interface the exploratory view of the clustering result. There are three visualization in this interface that shows the aspect (*root_type* and *day*) frequency and correlation. Regarding the frequency visualization, this view shows the relative frequency of aspects for the trajectories grouped together in a given node. Likewise, the correlation visualization is related to aspects of trajectories in the same node. For simplicity purpose, Figure 4 illustrates the correlation visualization placed in this interface. Therefore, the analyst can interact with the node and user filters to check some aspect behaviour among the grouped users.

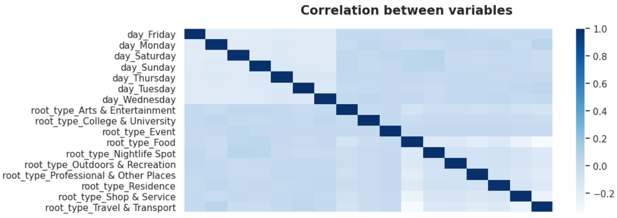
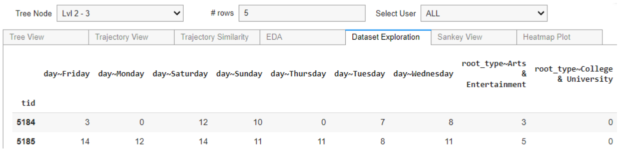


Figure 4: EDA view with correlation plot

## 2.5 Dataset exploration

The *dataset exploration* interface is a simple visualization of the frequency matrix. It is used to support the analyst on expecting the aspect frequency values using the filters node and number of rows. Figure 5 shows the aspect frequencies of five trajectories. The analyst can use the scroll bar to visualize other aspects and select the number of trajectories (*rows*) to display on the interface. Thus, such view is another resource to support the analyst on the task of checking the frequency occurrences.


Tree Node: Lvl 2 - 3 # rows: 5 Select User: ALL

| | day~Friday | day~Monday | day~Saturday | day~Sunday | day~Thursday | day~Tuesday | day~Wednesday | root_type~Arts & Entertainment | root_type~College & University |
|---|---|---|---|---|---|---|---|---|---|
| tid | | | | | | | | | |
| 5184 | 3 | 0 | 12 | 10 | 0 | 7 | 8 | 3 | 0 |
| 5185 | 14 | 12 | 14 | 11 | 11 | 8 | 11 | 5 | 0 |

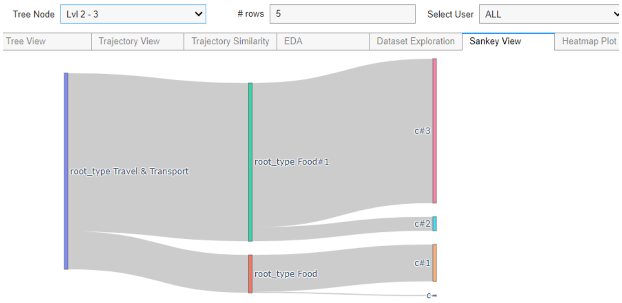
Figure 5: Dataset exploration for aspect frequency values

## 2.6 Sankey aspect relevance

The *Sankey Aspect Relevance* diagram present a unique view of the aspects role in the clustering trajectories. This visualization shows the selection aspect sequence, i.e. the clustering tree path. The aspect sequence starts with the most relevant aspect (root node) until the leaf nodes (clusters). Each column represents the tree level, while each bar represents an aspect in that level which its height indicates how many trajectories contain that aspect. Figure 6 illustrates the aspect flow starting with *Travel & Transport*, then passing by *Food* and finish at cluster level. Similar to *trajectory view*, the Sankey interface does not interact with any filter because it shows the overall clustering flow among the relevant aspects. This visualization can support the analyst on visualize the flow and interaction among the aspects.

## 2.7 Heatmap

The *heatmap* interface illustrates the frequency matrix into one plot. It represents each trajectory of a node as a row while the aspects (*day* and *root_type*) in the trajectory set as a column in the visualization. This view just interact with the node filter, but it is possible to extend the user filter to make it interact with this view. For instance, Figure 7 shows the heatmap visualization for node 3 at level 2 (259 trajectories). We note that this cluster mainly spread up, however, there is a gap in the middle which identifies a missing pattern in this set of trajectories. This gap contains the aspects *Arts & Entertainment*, *College & University* and *Event*, thereby, it



Tree Node Lvl 2 - 3

# rows 5

Select User ALL

**Figure 6: Sankey aspect view**

is a cluster that does not frequently visit places of this category. Thus, such interface is other resource that can support the analyst to identify patterns on trajectories.

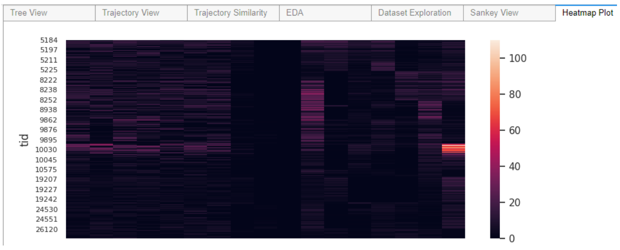


**Figure 7: Heatmap view**

## 3 CONCLUSION AND FUTURE WORK

We show MAT-CA which offers tools to MAT clustering analysis. We introduced in this work some functionalities of the python using widgets to build this tool. It is structured to deal with MATs, but it supports general multidimensional datasets. Besides, it offers different options to visualize the clustering result. We argue that MAT-CA can help the analyst on daily tasks with interactive visualizations and decreasing time with hard coding. MAT-CA is idealized for MATs, but as an open source work it can be extended for other needs. As future works, we want to extend this prototype with other trajectory clustering methods as well as other types of visualizations, and scale up it for a web-based platform.

## ACKNOWLEDGMENTS

This study was financed in part by CAPES - Finance Code 001. This work was also supported by CNPq, FAPESC - Project Match - co-financing of H2020 Projects - Grant 2018TR 1266, and by the MASTER project that has received funding from the European Union's Horizon 2020 research and innovation programme under the MarieSlodowska Curie grant agreement N.777695.

## REFERENCES

[1] Gennady Andrienko, Natalia Andrienko, Georg Fuchs, and Jose Manuel Cordero Garcia. 2017. Clustering trajectories by relevant parts for air traffic analysis. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 34–44.

[2] Mohd Yousuf Ansari, Amir Ahmad, Gopal Bhushan, et al. 2021. Spatiotemporal trajectory clustering: A clustering algorithm for spatiotemporal data. *Expert Systems with Applications* (2021).

[3] Muhammad Bilal, Shahid Masud, and Shahrukh Athar. 2012. FPGA design for statistics-inspired approximate sum-of-squared-error computation in multimedia applications. *IEEE Transactions on Circuits and Systems II: Express Briefs* (2012), 506–510.

[4] Yang Fan, Qing Xu, Yuejun Guo, and Sheng Liang. 2015. Visualization on agglomerative information bottleneck based trajectory clustering. In *2015 19th International Conference on Information Visualisation*. IEEE, 557–560.

[5] Carlos Andres Ferrero, Luis Otavio Alvares, and Vania Bogorny. 2016. Multiple aspect trajectory data analysis: research challenges and opportunities.. In *GeoInfo*. 56–67.

[6] Yuejun Guo, Qing Xu, and Mateu Sbert. 2018. IBVis: Interactive visual analytics for information bottleneck based trajectory clustering. *Entropy* 20, 3 (2018), 159.

[7] Jing He, Haonan Chen, Yijin Chen, Xinming Tang, and Yebin Zou. 2019. Variable-based spatiotemporal trajectory data visualization illustrated. *IEEE Access* 7 (2019), 143646–143672.

[8] Chih-Chieh Hung, Wen-Chih Peng, and Wang-Chien Lee. 2015. Clustering and aggregating clues of trajectories for mining trajectory patterns and routes. *The VLDB Journal* 24 (2015), 169–192.

[9] Dheeraj Kumar, Huayu Wu, Sutharshan Rajasegarar, Christopher Leckie, Shonali Krishnaswamy, and Marimuthu Palaniswami. 2018. Fast and scalable big data trajectory clustering for understanding urban mobility. *IEEE Transactions on Intelligent Transportation Systems* (2018), 3709–3722.

[10] Caihong Liu and Chonghui Guo. 2020. STCCD: Semantic trajectory clustering based on community detection in networks. *Expert Systems with Applications* 162 (2020), 113689.

[11] Yingchi Mao, Haishi Zhong, Hai Qi, Ping Ping, and Xiaofang Li. 2017. An adaptive trajectory clustering method based on grid and density in mobile pattern analysis. *Sensors* (2017), 2013.

[12] Ronaldo dos Santos Mello, Vania Bogorny, Luis Otavio Alvares, Luiz Henrique Zambom Santana, Carlos Andres Ferrero, Angelo Augusto Frozza, Geomar Andre Schreiner, and Chiara Renso. 2019. MASTER: A multiple aspect view on trajectories. *Transactions in GIS* 23, 4 (2019), 805–822.

[13] Fanrong Meng, Guan Yuan, Shaoqian Lv, Zhixiao Wang, and Shixiong Xia. 2019. An overview on trajectory outlier detection. *Artificial Intelligence Review* 52 (2019), 2437–2456.

[14] Brendan Morris and Mohan Trivedi. 2009. Learning trajectory patterns by clustering: Experimental studies and comparative evaluation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 312–319.

[15] Alireza Naghizadeh and Dimitris N Metaxas. 2020. Condensed silhouette: An optimized filtering process for cluster selection in K-means. *Procedia Computer Science* (2020), 205–214.

[16] Lucas May Petry, Carlos Andres Ferrero, Luis Otavio Alvares, Chiara Renso, and Vania Bogorny. 2019. Towards semantic-aware multiple-aspect trajectory similarity measuring. *Transactions in GIS* 23, 5 (2019), 960–975.

[17] Fabio Poiesi and Andrea Cavallaro. 2015. MTTV-An Interactive Trajectory Visualization and Analysis Tool.. In *IVAPP*. 157–162.

[18] Tarlis Tortelli Portela, Vania Bogorny, Anna Bernasconi, and Chiara Renso. 2022. AutoMATise: Multiple Aspect Trajectory Data Mining Tool Library. In *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*. Online, 282–285. https://doi.org/10.1109/MDM55031.2022.00060

[19] Yuri Santos, Ricardo Giuliani, Tarlis Portela, and Jônata Tyska. 2023. MAT-Tree: A Tree-based Method for Multiple Aspect Trajectory Clustering. (2023), to appear.

[20] Chunhua Tang, Meiyue Chen, Jiahuan Zhao, Tao Liu, Kang Liu, Huaran Yan, and Yingjie Xiao. 2021. A novel ship trajectory clustering method for Finding Overall and Local Features of Ship Trajectories. *Ocean Engineering* (2021), 110108.

[21] Iraklis Varlamis, Christos Sardianos, Vania Bogorny, Luis Otavio Alvares, Jônata Tyska Carvalho, Chiara Renso, Raffaele Perego, and John Violos. 2021. A novel similarity measure for multiple aspect trajectory clustering. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. ACM, 551–558.

[22] Xu Wang and Yusheng Xu. 2019. An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. In *IOP Conference Series: Materials Science and Engineering*. IOP Publishing.

[23] Yulong Wang, Kun Qin, Yixiang Chen, and Pengxiang Zhao. 2018. Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi GPS data. *ISPRS International Journal of Geo-Information* (2018).

[24] Junwei Xiao, Jianfeng Lu, and Xiangyu Li. 2017. Davies Bouldin Index based hierarchical initialization K-means. *Intelligent Data Analysis* (2017), 1327–1338.

[25] Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. 2014. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 1 (2014), 129–142.

[26] Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. 2014. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 1 (2014), 129–142.