EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems Research Article EAI.EU

Histogram-based Feature Extraction for GPS Trajectory Clustering

Chi Nguyen[1], Thao Dinh[2], Van-Hau Nguyen[3], Nhat Phuong Tran[4] and Anh Le[1,*]

[1] Ho Chi Minh City, University of Transport, Vietnam
[2] Department of Information Technology and Resources and Environment Data, Vietnam
[3] Hung Yen University of Technology and Education, Vietnam
[4] The Institute of Electronics, Communications and Information Technology (ECIT), Queen's University Belfast, UK

Abstract

Clustering trajectories from GPS data is a crucial task for developing applications in intelligent transportation systems. Most existing approaches perform clustering on raw data consisting of series of GPS positions of moving objects over time. Such approaches are not suitable for classifying moving behaviours of vehicles, e.g., how to distinguish between a trajectory of a taxi and a trajectory of a private car. In this paper, we focus on the problem of clustering trajectories of vehicles having the same moving behaviours. Our approach is based on histogram-based feature extraction to model moving behaviours of objects and utilizes traditional clustering algorithms to group trajectories. We perform experiments on real datasets and obtain better results than existing approaches.

Keywords: trajectory clustering, histogram, data clustering, GPS.

Received on 12 October 2019, accepted on 02 January 2020, published on 17 January 2020

doi: 10.4108/eai.13-7-2018.162796

*Corresponding author. Email:anhlvq@gmail.com

1. Introduction

With the ever-increasing number of smartphones and mobile devices equipped with Global Positioning Systems (GPS) receivers, large amounts of spatio-temporal data can be collected from moving objects, e.g., vehicles or travellers. Such massive quantity of data has led to a rise in the number of data mining tasks aiming to analyse and discover useful information for real life applications [1]. One of the tasks is clustering GPS trajectory data to find frequent traffic flows in urban traffic networks toward developing applications in intelligent transportation systems (ITS).

Basically, GPS trajectory data is collected for a moving object to track its positions over time. Such data is also called GPS logs. GPS trajectory data can be considered a kind of big data and often contains noisy, missing, and incorrect values inside [2][3]. Moreover, there is no standard for GPS logs in terms of data format. Therefore, mining from such data is a big challenge for data scientists.

Most of existing approaches to tackle the problem of finding similar trajectories is to apply traditional clustering algorithms on raw trajectory data. Such kinds of approaches miss similar partitions shared by trajectories which belong to different groups [4]. Moreover, clustering on raw trajectory data cannot detect groups of objects having the same moving behaviours, e.g., moving of cars vs moving of motorbikes.

In this paper, we proposed an approach to cluster trajectory data, which can detect groups of vehicles having similar moving behaviours. The basic idea is based on extracting features which describe moving behaviours of vehicles from raw trajectory data. In particular, we employ histogram-based features to model how a vehicle moves on the urban traffic networks, e.g., about speed, acceleration, or frequent visiting places. We use a real-world data source for the experiments and obtain good results.

Before describing the proposed approach in details, we first review the related work and then define some concepts to formulate the problem.

2. Related Work

In general, traditional clustering approaches can be directly applied to find groups of trajectories. These approaches heavily relied on a predefined distance measure between two trajectories [2]. There are many different ways to define such a measure, e.g., using Euclidean distance, Edit distance, or Dynamic Time Wrapping. However, trajectory data have been found to be inaccurate, highly sensitive of sampling methods, and have low robustness for the noisy data [5]. Using raw trajectory data as sequence of sampled GPS points to compute the distance might lead to obtain unexpected result or sometime meaningless in clustering tasks. In the context of detecting moving behaviours, such kinds of approaches might not be helpful.

Another direction to tackle the problem is first mapping raw trajectory data to some feature space, and then defining a distance measure on that [2][6][7][8][9]. The most similar idea to our approach is presented by D.Yao et al. [10], where the authors using velocity of vehicles to extract moving behaviour features. We observe that such kind of features cannot work well for cases of mixed traffic flows that widely exist in developing countries.

Different from the above approaches, we propose an approach that based on histogram-based feature extraction to represent raw trajectory data. Clustering is performed on histogram-based feature space. In the following sections, we first define necessary concepts, then show the proposed framework to detect similar moving behaviours from GPS trajectory data.

3. Basic Concepts and Notations

3.1. Modelling GPS Trajectory Data

Basically, the data obtained from GPS-enabled devices exists in the form of GPS log. As shown in the Fig.1, the data consists of a series of records, where each record describes the position and the status of a vehicle at a particular timestamp. This kind of data format is considered raw data in this paper and will be formulated to the more descriptive form to be mined.

```
"EVENT_ID","ACCOUNT_ID","TIMESTAMP","STATUS_CODE","LATITUDE","LONGITUDE","
625653769,0,19-JUL-14 07.48.50.000000000 AM,16417,16.06716,108.2431866667,
625976041,0,19-JUL-14 07.49.04.000000000 AM,16416,16.0678833333,108.24325
625976086,0,19-JUL-14 07.49.05.000000000 AM,16416,16.06795,108.2432333333,
625976385,0,19-JUL-14 07.49.20.000000000 AM,16416,16.0687633333,108.24333
625976711,0,19-JUL-14 07.49.35.000000000 AM,16416,16.0688183333,108.24368
625977009,0,19-JUL-14 07.49.50.000000000 AM,16416,16.0687283333,108.24406
625977357,0,19-JUL-14 07.50.35.000000000 AM,16417,16.06888,108.2450116667,
625977342,0,19-JUL-14 07.50.35.000000000 AM,16417,16.06888,108.2450116667,
625978335,0,19-JUL-14 07.50.50.000000000 AM,16416,16.0691916667,108.24529
625978644,0,19-JUL-14 07.51.05.000000000 AM,16416,16.07023,108.2453033333,
625978687,0,19-JUL-14 07.51.06.000000000 AM,16416,16.0703066667,108.2453,,
625979001,0,19-JUL-14 07.51.19.000000000 AM,16416,16.0715933333,108.24527
625979323,0,19-JUL-14 07.51.34.000000000 AM,16416,16.0733016667,108.24521
625979646.0.19-JUL-14 07.51.49.000000000 AM.16416.16.0750633333.108.24521
```

Fig.1. An example of GPS Log

In this paper, we use the following concepts to formulate the problem:

- **GPS point**: a GPS point is represented by a tuple *<id, lat, lon, time>*, where: *id* is the identifier of the moving object; *lat* is the latitude'; *lon* is the longitude; and *time* is the timestamp that the GPS point is reported.
- **GPS Log**: a GPS log is a dataset that consists of GPS points.
- **GPS Trajectory**: Let S be a sequence of GPS points, in the form of $p_1 \rightarrow p_2 \rightarrow \ldots \rightarrow p_n$, where $p_1, p_2, \ldots p_n$ are GPS points; Δ_t be the predefined time interval threshold; Δ_d be the predefined distance threshold, the sequence S is called a GPS trajectory if and only if all the following conditions are satisfied:
  - *i.* $p_i.id = p_j.id, \forall i, j \in \{1, 2, \ldots, n\}$
  - *ii.* $p_i.time - p_{i-1}.time \leq \Delta_t, \forall i \in \{2, \ldots, n\}$
  - *iii.* $||p_i - p_{i-1}|| \leq \Delta_d, \forall i \in \{2, \ldots, n\}$

Fig.2 shows an example of a GPS trajectory as a sequence of GPS points that are not interrupted in both space and time.

Fig.2. An example of a GPS Trajectory

3.2. Histogram-based Feature Extraction from GPS Trajectories

In this work, we consider each GPS trajectory as an object and perform clustering on a set of GPS trajectories extracted from GPS Log. Before doing that, we first map GPS trajectories to a feature space. We observed that moving behaviours can be described by using histogram of frequently visited places. In this work, we use a grid NxM to divide space into equal size cells. A GPS trajectory can be represented by a matrix *NxM*, where the value at row *i* and column *j* of the matrix stands for the number of GPS Points belonging to the cell *[i, j]*. Clearly, the higher the value is, the longer time the moving object stays in the corresponding cell. For example, a bus often stays at bus stations only, a taxi might stay at any places on the road network.

Note that, a cell having a high value has the same meaning as a stay-point in existing works. However, while existing works compute stay-points from the whole dataset, stay-points in this work is computed for each individual GPS trajectory to compose the feature for that trajectory.

Fig.3 shows an example of the transformation from a GPS trajectory to histogram-based feature. The grid size in this example is 50x50. One can see the darker the cells are, the more frequently the vehicle stays.

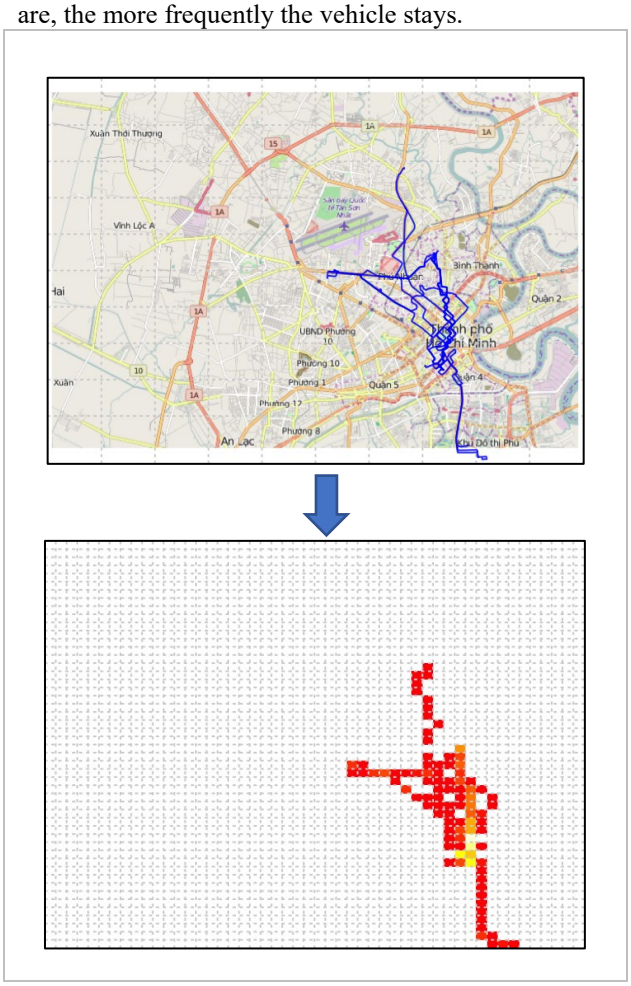


Fig.3. An illustration of histogram-based feature extracted from a trajectory. The grid size is 50x50.

4. Our Framework to Detect Similar Moving Behaviours

In this section, we describe our framework to detect groups of objects that have similar moving behaviours. This framework contains three steps. The first step is the data pre-processing step where data cleaning is performed to remove noises and incorrect data. GPS trajectories are then extracted from the raw data by using predefined pair of thresholds as described in Section 3.1.

The second step is feature extraction, whereas each GPS trajectory is transformed to a matrix describing 2D histogram of GPS points in the trajectory. In this step, PCA method can be applied to reduce the dimension of the feature space.

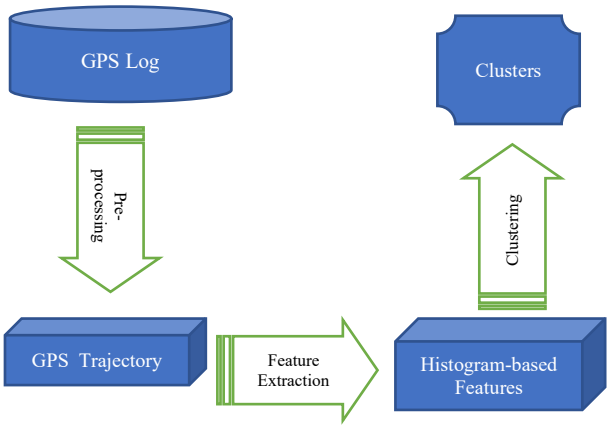


Fig.4. The Framework to detect similar moving behaviours.

Finally, one can apply any clustering algorithm to group the trajectories in the third step. In our experiment described later on, we use DBSCAN [11], [12] for the clustering task.

5. Experimental Evaluation

5.1. Datasets and Experimental Setup

For real datasets, we obtain GPS Log from a company providing vehicle tracking services, called OTS. In this dataset, we have 411 different vehicles of the following

type: bus, truck, taxi, and private car, moving around Ho Chi Minh City, Vietnam. The data is collected for one week from June 01, 2015 to June 07, 2015.

5.2. Data Pre-processing Step

The real datasets are very noisy and consists of incorrect data. We perform pre-processing step as desrcibed in Section 4. In our experiments, we use 30 minutes as a threshold value for the time interval and 1 km as a threshold value for the distance.

Fig. 5 shows orgininal data obtained from OTS company that containing noisy and incorrect data. Fig. 6 shows that data has been cleaned up. One can see the pre-processing step is very important to the mining step later on.

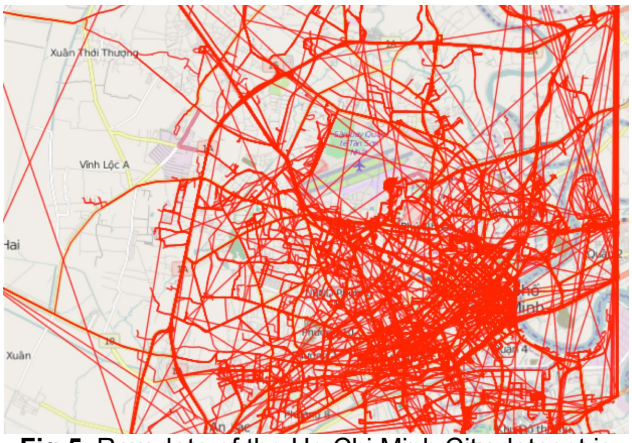


Fig.5. Raw data of the Ho Chi Minh City dataset is very noisy.

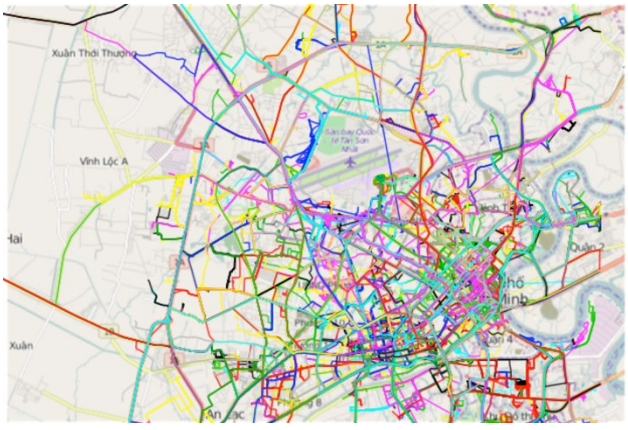


Fig.6. The Dataset after the pre-processing step.

5.3. Feature Extraction

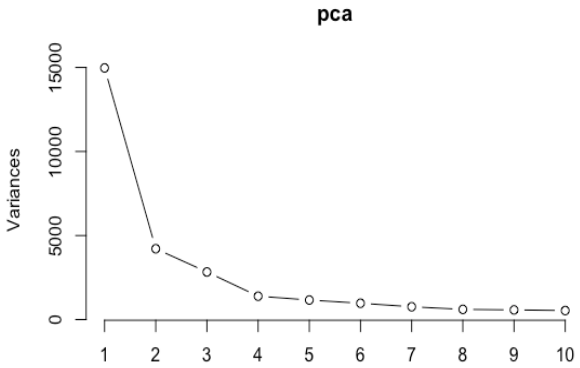
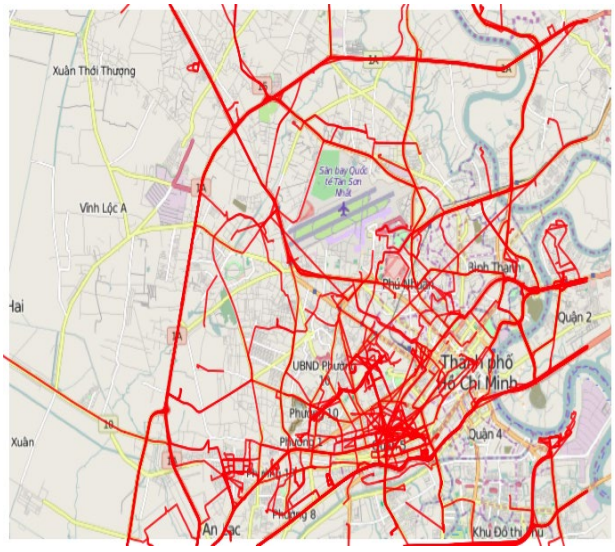


Fig.7. Apply PCA for the dimentionality reduction step.

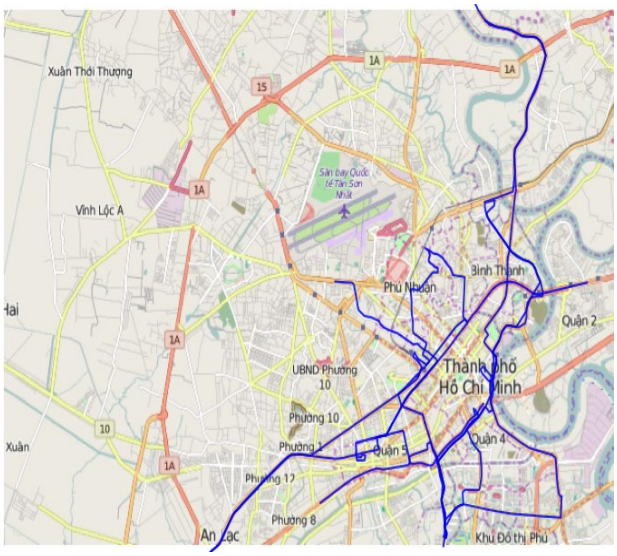
We perform the histogram-based feature extraction step and then apply PCA method to reduce the dimensionality. In our experiments, we can reduce the dimensionality to 6, as shown in Fig. 7.

5.3. Clustering Performance and Results

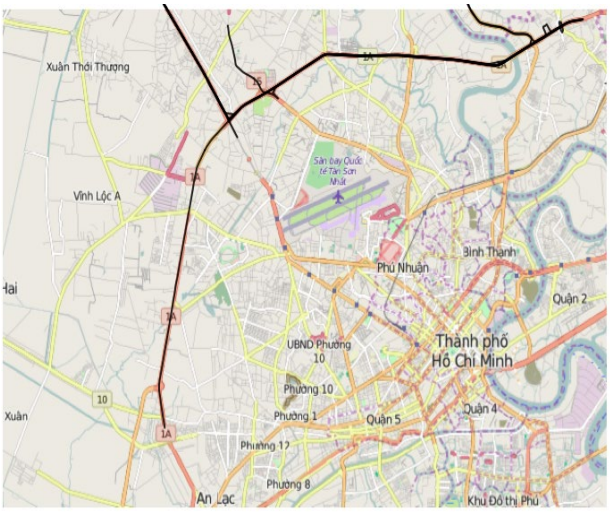
We employ DBSCAN algorithm [8,9] to perform clustering task on the feature space. We obtain 4 clusters that are shown in Fig.8 and Fig.9.



Cluster 1: Trajectories of private cars



Cluster 2: Trajectories of public buses



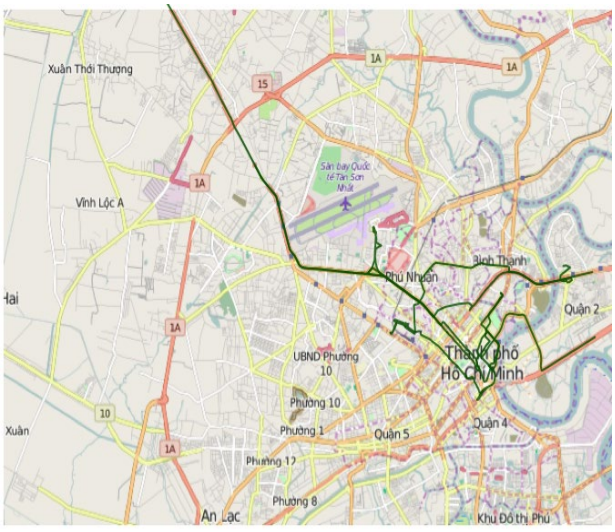
Cluster 4: Trajectories of vehicles passing through Ho Chi Minh City

Fig.8. Clustering results obtained from Ho Chi Minh City Dataset: difference between private and public transport.

Fig.9. Clustering results obtained from Ho Chi Minh City Dataset: difference between transit buses and long-distance buses.

After take a look closer inside the dataset, these clusters can be explained as follows:

- Cluster 1 contains trajectories of vehicles driving on almost roads of the road network. They are private cars or taxies.
- Cluster 2 contains trajectories of buses traveling with fixed routes.
- Cluster 3 contains long-distance bus travel to Ho Chi Minh City.
- Cluster 4 contains trajectories of vehicles passing through Ho Chi Minh City.



Cluster 3: Trajectories of transit buses

The results from the experiments shows that the histogram-based feature extraction can be employed for discovering meaningful clusters from raw GPS trajectories.

We compare our clustering results to the most similar approach proposed by D.Yao [10]. We use the same framework, except the feature extraction steps. The clustering results cannot clearly be interpreted. For example, Figure 10 shows a cluster that contains mixing types of vehicles and it is hard to identify groups of vehicles having the same moving behaviours. This can be explained by the fact that velocity is not a good feature to identify the type of a moving object in mixed traffic flows that widely exist in developing countries.

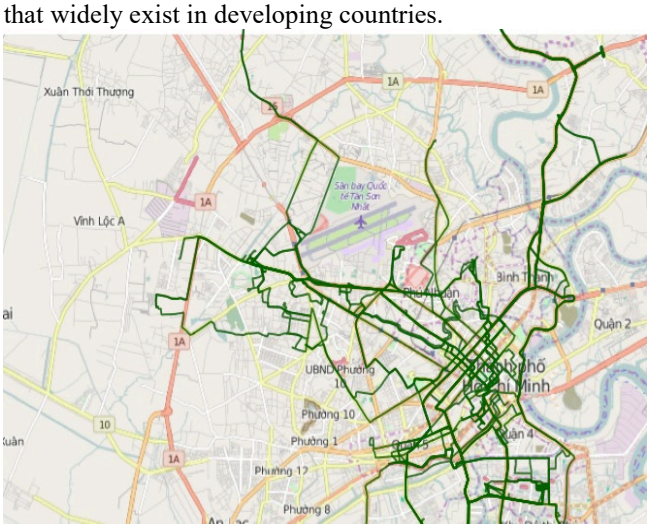


Fig.10. Clustering based on velocity features from Ho Chi Minh City Dataset: each cluster contains mixing types of vehicles.

6. Conclusions and Ongoing Work

In this work, we present an approach to the problem of trajectory clustering where each GPS trajectory is transformed into histogram-based features, as a form of 2 dimensional array. We also apply PCA method to reduce the dimensionality before a clustering task is performed. We compare our experimental results on real datasets to other feature extraction approaches and the histogram-based feature extraction is shown to be a good feature extraction approach to detect moving behaviours of vehicles.

Nowadays, trajectory data is collected continuously in a streaming manner[13]. This work can be extended to adapt such a streaming data. In this case, the framework is able to incrementally compute the histogram-based features to keep the data up-to-date. This direction will be considered in our future work.

Acknowledgements.

This work is supported by research grant No. KH1902 from HCMC University of Transport.

References

[1] W. Tang, D. Pi, and Y. He, "A density-based clustering algorithm with sampling for travel behavior analysis," *Lecture Notes in Computer Science*, vol. 9937 LNCS. pp. 231–239, 2016.

[2] S. Atev, G. Miller, and N. P. Papanikolopoulos, "Clustering of vehicle trajectories," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 647–657, 2010.

[3] X. Zhou, F. Miao, H. Ma, H. Zhang, and H. Gong, "A trajectory regression clustering technique combining a novel fuzzy C-means clustering algorithm with the least squares method," *ISPRS Int. J. Geo-Information*, vol. 7, no. 5, pp. 9–16, 2018.

[4] J. H. Jae-Gil Lee and Kyu-Young Whang, "Trajectory Clustering: A Partition-and-Group Framework," in *SIGMOD'07*, 2007, pp. 593–604.

[5] Y. U. Zheng, "Trajectory Data Mining : An Overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, pp. 1–41, 2015.

[6] X. Zhou *et al.*, "An Automatic K-Means Clustering Algorithm of GPS Data Combining a Novel Niche Genetic Algorithm with Noise and Density," *ISPRS Int. J. Geo-Information*, vol. 6, no. 12, p. 392, 2017.

[7] Z. Ding, B. Yang, R. H. Güting, and Y. Li, "Network-Matched Trajectory-Based Moving-Object Database : Models and Applications," pp. 1–11, 2015.

[8] Z. Chen, H. T. Shen, and X. Zhou, "Discovering popular routes from trajectories," *Proc. - Int. Conf. Data Eng.*, vol. 4, no. c, pp. 900–911, 2011.

[9] S. Wang, Z. Bao, J. S. Culpepper, T. Sellis, and X. Qin, "Fast large-scale trajectory clustering," *Pvldb*, vol. 13, no. 1, pp. 29–42, 2019.

[10] D. Yao, C. Zhang, Z. Zhu, J. Huang, and J. Bi, "Trajectory Clustering via Deep Representation Learning," *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017.

[11] S. T. Mai, I. Assent, and M. Storgaard, "AnyDBC: An efficient anytime density-based clustering algorithm for very large complex datasets," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[12] S. T. Mai, I. Assent, J. Jacobsen, and M. S. Dieu, "Anytime parallel density-based clustering," *Data Min. Knowl. Discov.*, 2018.

[13] J. Mao, Q. Song, C. Jin, Z. Zhang, and A. Zhou, "Online clustering of streaming trajectories," *Front. Comput. Sci.*, vol. 12, no. 2, pp. 245–263, 2018.