Offline Trajectory Analytics (Short Description)

Panagiotis Tampakis, Stylianos Sideridis, Panagiotis Nikitopoulos, Nikos Pelekis, Christos Doulkeridis and Yannis Theodoridis

1 Motivation

Concerning the analysis of mobility data, mobility data analytics aim to describe the mobility of objects, to extract valuable knowledge by revealing motion behaviors or patterns, to predict future mobility behaviors or trends and in general, to generate various perspectives out of data, useful for many other scientific fields. An important operation when trying to extract knowledge out of mobility data is cluster analysis, which aims at identifying clusters of moving objects, as well as detecting moving objects that demonstrate abnormal behavior and can be considered as outliers. Several efforts try to identify patterns that are valid for the entire lifespan of the moving objects [2, 5, 1]. However, discovering clusters of complete trajectories can overlook significant patterns that might exist only for some portions of their lifespan, which motivates us to try deal with the problem of subtrajectory clustering. What is even more challenging is to trying to deal with this problem in the Big Data era, which calls for parallel and distributed algorithms in order to address the scalability requirements. In this context, one challenge is how to partition the data in such a way so that each node can perform its computation independently, thus minimizing the communication cost between nodes, which is a cost that can turn out to be a serious bottleneck. Another challenge, related to partitioning, is how to achieve load balancing, in order to balance the load fairly between the different nodes. Yet another challenge is to minimize the iterations of data processing, which are typically required in clustering algorithms.

In geospatial analysis, a hotspot is a geographic area that contains unusually high concentration of activities (e.g. moving objects). Trajectory hotspot analysis is a special case of geospatial analysis, which discovers spatio-temporal regions having

Panagiotis Tampakis, Stylianos Sideridis, Panagiotis Nikitopoulos, Nikos Pelekis, Christos Doulkeridis and Yannis Theodoridis
University of Piraeus, Karaoli & Dimitriou St. 80

e-mail: {ptampak,ssider,nikp,npelekis,cdoulk,ytheod}@unipi.gr

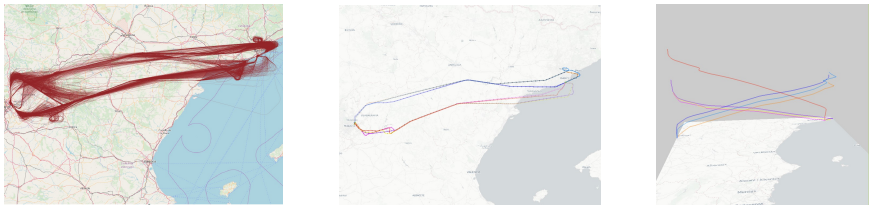
high concentration of moving trajectories. Motivated by the need for big data analytics over trajectories of vessels, we focus on discovering *trajectory hotspots* in the maritime domain, as this relates to various challenging use-case scenarios. For example, having a predefined set of regions of interest, for which there is a priori knowledge about occurring activities in them, it is very useful to be able to analyze (a) the intensity of the fishing activity (i.e., fishing pressure) of the areas, or (b) to quantify the environmental fingerprint by the passage of a particular type of vessels from the areas. Similar cases exist in all mobility domains. Thus, the effective discovery of such diverse types of hot spots is of critical importance for our ability to comprehend the various domains of mobility.

Inference of the underlying network given a large number of moving traces, both in aviation and maritime domain is a challenging task that we try to address. The goal is to discover the directed graph of transitions, i.e., the set V of vertices and the set E of edges that form the routes network. Additionally, enriched information has to be taken into account in order to produce an enriched graph with contextual information. Domain experts may benefit a lot from such additional information. For example, one can then easily produce analytics of trajectories based on specific weather conditions and reveal how these conditions affecting or not the paths followed. Moreover, flight plans or predefined sea routes can be compared with real paths followed by ships or planes and the domain expert would be able to identify and explain the reason more easily.

2 Distributed (Sub)Trajectory Clustering

Our approach to distributed subtrajectory clustering [7] splits the problem in three steps. The first step is to retrieve, in a distributed way, for each trajectory $r \in D$, all the moving objects, with their respective portion of movement, that moved close enough in space and time with r, for at least some time duration. This is a well-defined problem in the literature of mobility data management, known as *subtrajectory join* [6] (the case of self-join). The subtrajectory join will return for each pair of (sub)trajectories, all the common subsequencies that have at least some time duration, which are actually candidates for the longest common subsequence. The second step takes as input the result of the first step and aims at segmenting, in a distributed fashion, each $r \in D$ into a set of subtrajectories D'. In our case, the way that a trajectory is segmented into subtrajectories is neighbourhood-aware, meaning that a trajectory will be segmented every time its neighbourhood changes significantly. Finally, the third step takes as input D' and the goal is to identify, in a distributed manner, clusters (whose cardinality is unknown) of similar subtrajectories and at the same time identify subtrajectories that are significantly dissimilar from the others (outliers). To illustrate the quality of the results we employed the IFS (April 2016) (Figure 1(a)). Figure 1(b), depicts the results of the subtrajectory clustering algorithm and Figure 1(d) the corresponding space-time cube. The algorithm identified 6 clusters, 3 clusters from Madrid to Barcelona and 3 clusters from Barcelona

to Madrid. Moreover, an outlier was detected, which is not something common in aviation data.



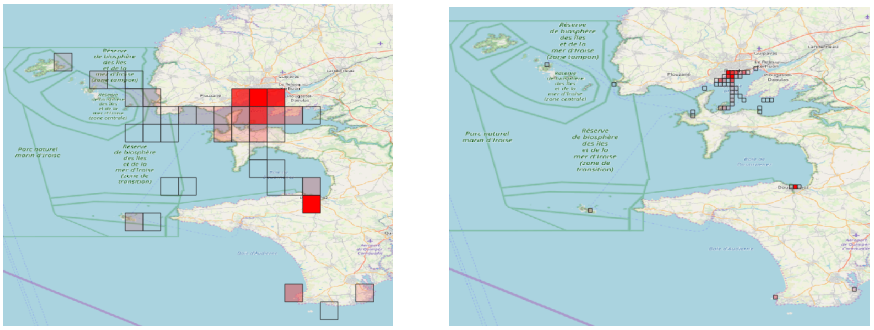
(a) (b) (c)

Fig. 1 (a) Raw data, (b) cluster representatives (6 clusters discovered), (c) cluster representatives space time cube

3 Distributed Hotspot Analysis

The problem of discovering trajectory hotspots over distributed sets of data is studied in [3], where two algorithms are proposed, namely THS and $aTHS$ for efficiently discovering trajectory hotspots in parallel. This approach is based on spatio-temporal partitioning of the 3D data space in cells. Accordingly, it tries to identify cells that constitute hotspots, i.e., not only do they have high density, but also that the density values are statistically significant. To this end, it employs the Getis-Ord statistic [4], a popular metric for hotspot analysis, which produces z-scores. The Getis-Ord statistic uses attribute values to provide z-scores for each cell. The attribute values represent the density of moving trajectories inside a specific cell. A formal definition of a cell's attribute value is provided in [3], which is based on the duration that an object is moving in the cell divided by the total lifespan of that cell. This definition implies that an attribute value is increased by having more vessels moving for longer duration in the cells of interest. The Getis-Ord statistic calculates a z-score for a cell by aggregating its attribute value with the attribute values of all the other cells of interest. A weight factor is used which determines the effect of a cell's attribute value to the z-score calculation of a neighboring cell. It represents the score influence between neighboring cells: a cell needs to have a neighborhood of high attribute values to be considered as hotspot. The goal is to have the influence of neighboring cells to be decreasing with increased spatio-temporal distance. Thus the study employs a weight function that decreases exponentially with increasing distance. $aTHS$ algorithm is able to constraint the influence distance, since for longer distances the influence becomes increasingly low. Hence, $aTHS$ approximates the final result, by providing analytical error bounds.

Fig. 2 demonstrates the top-50 hot-spots discovered by THS algorithm for a data set covering the Brest area, based on a user-defined grid. Each hot-spot is a region



(a) (b)

Fig. 2 Hot-spots on (a) large regions and (b) small regions.

defined by a rectangular cell, which is part of the grid provided for the entire data set. The size of the cells in Fig. 2(a) is 0.05 degrees in both longitude and latitude dimensions, while in Fig. 2(b) the cell size is configured to be 0.01 degrees

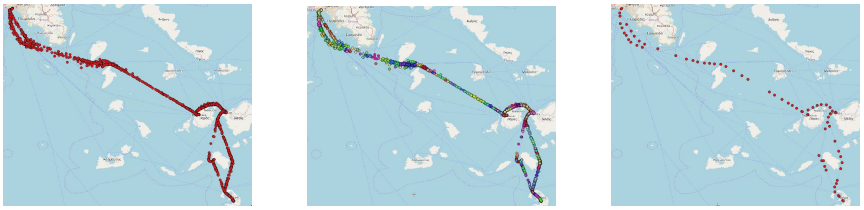
In the experimental section of [3], the proposed approach is evaluated over a real set of data containing surveillance information from the maritime domain. The data was collected over a period of three years, consisting of individual trajectories for vessels moving in the Eastern Mediterranean Sea. The efficiency of the proposed approach is affected mainly by the neighborhood influence calculation step. However, due to $aTHS$ ability to constraint the influence distance, the proposed approach calculates an approximate result in reasonable time, while providing error guarantees.

4 Distributed Data-enriched Mobility Networks

We follow an approach where first the vertices are discovered and then edges connecting these vertices are inferred from the trajectories. The input to the process comprises of a set of enriched trajectories. An enriched trajectory is modeled as a sequence of time-stamped enriched points. The output of the process is a semantic aware mobility network modeled as a directed graph $G = (V, E)$, where the vertices V correspond to semantic nodes and the edges E correspond to the discovered paths between semantic nodes.

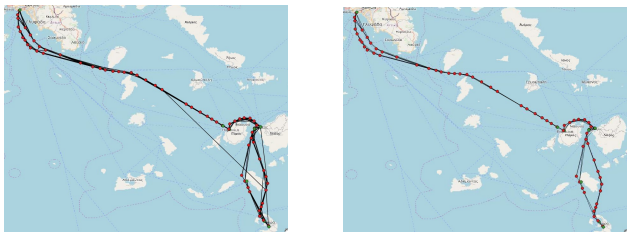
The process comprises two main steps, *Enriched-nodes-extraction* and *Enriched-paths-discovery*. Figure 3 illustrates the Enriched nodes extraction step, while the Enriched paths discovery step is depicted in Figure 4.

Data-driven contextually aware inference of transportation network map is the main outcome of the proposed methodology. From the qualitative evaluation, it can be concluded that first, the higher the weight of the edges the higher the compactness in the representation of the dataset with this data-enriched network structure. Second, the network extracted from synopses is more or less the same as the one



(a) (b) (c)

Fig. 3 Overview of network nodes extraction step in maritime domain: (a) all enriched points from input, (b) enriched points clustered spatially to candidate nodes (c) semantic nodes extraction.



(a) (b)

Fig. 4 Overview of network paths discovery step in maritime domain: (a) all paths found (b) only edges with more than σ weight are kept.

produced by the raw data. The advantage of this is that not only we may extract the network by processing much less data, but more importantly, we gain from the contextual characterization of the synopses to attach semantics to the vertices of the network. The algorithms proposed provide contextually enhanced spatial graphs, which can successfully be utilized to support online location and trajectory prediction/forecasting scenarios. Moreover, the methodology is able to produce networks of high accuracy, which closely resemble the structure and topology of the underlying ground truth networks.

5 Contributions

Our main contributions, concerning the *Distributed Subtrajectory Clustering* problem, we formally define the problem of *Distributed Subtrajectory Clustering* (DSC) and propose two neighborhood-aware trajectory segmentation algorithms, which are tailored to DSC problem, covering different application requirements. Further, we design an efficient and scalable solution for the problem of *Distributed Subtrajectory Clustering*. Our experimental study, demonstrates the merits of our solution.

Regarding the problem of distributed trajectory hotspot analysis, we define the problem of trajectory hotspot analysis, as a special case of geospatial hotspot anal-

ysis, appropriately tailored to become meaningful for trajectories, rather than plain points. We presents two parallel algorithms for efficiently discovering trajectory hotspots in parallel over distributed sets of trajectory data. The first algorithm calculates the exact result, while the second improves the efficiency by providing an approximate result, limited by error bounds. Finally, we provide an efficiency and scalability emprical evaluation, by experimenting with real vessel trajectory data.

Concerning the distributed data-enriched mobility networks problem, we introduce a novel contextually aware network construction methodology that operates on annotated and durative critical points (synopses) which accurately represent the real motion of the vessels and aircraft. This contextually aware approach enables to combine new sea portions or flight routes due to incidents and updates. During the above process, we introduce a parameter less algorithm around critical points based on semantic similarity, which allows to create semantic nodes based on the available data by using sets of trajectories that belong to the same critical points category. We present a detailed validation study of our method, using real-world vessels and aircraft tracking data, which demonstrates the efficiency of the proposed method.

References

1. Deng, Z., Hu, Y., Zhu, M., Huang, X., Du, B.: A scalable and fast OPTICS for clustering trajectory big data. Cluster Computing **18**(2), 549–562 (2015)
2. Nanni, M., Pedreschi, D.: Time-focused clustering of trajectories of moving objects. J. Intell. Inf. Syst. **27**(3), 267–289 (2006)
3. Nikitopoulos, P., Paraskevopoulos, A., Doulkeridis, C., Pelekis, N., Theodoridis, Y.: Hot spot analysis over big trajectory data. In: IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018, pp. 761–770 (2018). DOI 10.1109/BigData.2018.8622376. URL https://doi.org/10.1109/BigData.2018.8622376
4. Ord, J.K., Getis, A.: Local spatial autocorrelation statistics: Distributional issues and an application. Geographical Analysis **27**(4), 286–306 (1995)
5. Pelekis, N., Kopanakis, I., Kotsifakos, E.E., Frentzos, E., Theodoridis, Y.: Clustering uncertain trajectories. Knowl. Inf. Syst. **28**(1), 117–147 (2011)
6. Tampakis, P., Doulkeridis, C., Pelekis, N., Theodoridis, Y.: Distributed subtrajectory join on massive datasets. ACM Trans. Spatial Algorithms Syst. **6**(2) (2019). URL https://doi.org/10.1145/3373642
7. Tampakis, P., Pelekis, N., Doulkeridis, C., Theodoridis, Y.: Scalable distributed subtrajectory clustering. In: IEEE BigData 2019, pp. 950–959 (2019)