Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9540944, 11 pages
https://doi.org/10.1155/2022/9540944

Research Article

Application of Artificial Intelligence in an Unsupervised Algorithm for Trajectory Segmentation Based on Multiple Motion Features

Wenjin Xu and Shaokang Dong

Department of Information Science Technology, Qingdao University of Science and Technology, No. 99 Songling Road, Laoshan District, Qingdao, China

Correspondence should be addressed to Shaokang Dong; 4019110002@mails.qust.edu.cn

Received 18 November 2021; Revised 27 November 2021; Accepted 1 December 2021; Published 6 January 2022

Academic Editor: Narasimhan Venkateswaran

With the development of the wireless network, location-based services (e.g., the place of interest recommendation) play a crucial role in daily life. However, the data acquired is noisy, massive, it is difficult to mine it by artificial intelligence algorithm. One of the fundamental problems of trajectory knowledge discovery is trajectory segmentation. Reasonable segmentation can reduce computing resources and improvement of storage effectiveness. In this work, we propose an unsupervised algorithm for trajectory segmentation based on multiple motion features (TS-MF). The proposed algorithm consists of two steps: segmentation and mergence. The segmentation part uses the Pearson coefficient to measure the similarity of adjacent trajectory points and extract the segmentation points from a global perspective. The merging part optimizes the minimum description length (MDL) value by merging local sub-trajectories, which can avoid excessive segmentation and improve the accuracy of trajectory segmentation. To demonstrate the effectiveness of the proposed algorithm, experiments are conducted on two real datasets. Evaluations of the algorithm's performance in comparison with the state-of-the-art indicate the proposed method achieves the highest harmonic average of purity and coverage.

1. Introduction

With the rapid development of location technology (such as GPS, Beidou System, AIS), it is becoming easier to get trajectory data of moving objects, including time, location, speed, acceleration, and heading. The analysis on trajectory data can provide a lot of valuable information for applications based on location data, such as traffic pattern detection [1], fishing detection [2, 3], animal migration behavior detection [4–6], human behavior patterns recognition [7], and hurricane trajectory prediction.

The preprocessing step of trajectory data mining includes noise cleaning, segmentation, stop points detection, compression, and map matching [8]. And trajectory segmentation is one of the most basic tasks, which is to partition the trajectory into disjoint parts. The motion features of each part are uniform, and the two adjacent parts represent different motion modes. Segmentation reduces computational complexity and allows us to mine richer knowledge, which exceeds the knowledge we learn from the entire trajectory. Furthermore, accurate segmentation methods can provide higher-quality features for further analysis of the behavior of moving objects.

In recent years, the proposed trajectory segmentation algorithms can be classified the supervised [9–14], unsupervised [15–27], and semisupervised [28].

The trajectory segmentation algorithm as aforementioned can solve most of the problems in the preprocessing

of trajectory data mining, but there are still the following challenges:

(1) At present, most of the supervised trajectory segmentation algorithms, such as SPD [11], Warped K-means [10], and WS-II [9], required labeled data or prior information such as time threshold, speed threshold, and the number of trajectory segments.

(2) Semisupervised trajectory segmentation algorithm (e.g., RGRASP-SemTS) uses a combination of both labeled and unlabeled data to segment. However, the majority of trajectory datasets do not contain the labeled data.

(3) The unsupervised trajectory segmentation algorithm does not require labeled data. But the existing unsupervised segmentation algorithms use greedy algorithms with high time complexity, resulting in uselessness which causes it is not suitable for large trajectory data.

To overcome these challenges, we propose an unsupervised algorithm for trajectory segmentation based on multiple motion features (TS-MF). The algorithm includes two steps: segmentation and mergence. First, to maximize the homogeneity of the subtrajectories, the segmentation part uses the Pearson coefficient to measure the similarity of trajectories. Furthermore, to avoid local oversegmentation, mergence part merges the subtrajectory by minimizing the cost function. Finally, we verify the proposed algorithm in two trajectory datasets of two different domains.

The main contributions of this article are as follows:

(1) The study proposes a segmentation method based on the Pearson coefficient. First, the Pearson coefficient is employed to measure the similarity according to the speed, acceleration, differential position, angle, and other movement features of the two trajectory points. Then, the trajectory is segmented from a global perspective.

(2) Considering the local oversegmentation of trajectories, we propose a merging method, which merges trajectories by minimizing cost function value.

(3) Fusion of the segmentation and merging method proposes an unsupervised algorithm for trajectory segmentation based on multiple motion features (TS-MF).

(4) The time complexity of our proposed algorithm is $O(n)$, which is suitable for the segmentation of large trajectory datasets.

The rest of this article is organized as follows: Section 2 gives the related works. Section 3 introduces the proposed trajectory segmentation algorithm. In Section 4, we verify the feasibility of our algorithm on two actual datasets. Finally, Section 5 gives our conclusions and future work.

2. Literature Review

In the past few years, scholars have published lots of papers related to trajectory segmentation. In this section, we mainly summarize most of the trajectory segmentation methods.

The supervised trajectory segmentation algorithm requires label data and heuristic rules such as the time threshold, speed threshold, density threshold, and angle threshold to segment trajectory. Mohammad et al. proposed a segmentation algorithm named WS-II [9], and it requires the labeled data. But the majority of trajectory datasets do not contain such information. Zheng et al. proposed a staying point detection (SPD) algorithm to segment trajectory [11]. SPD suppose that there is a stay point between two adjacent motion modes and uses the distance threshold δd and the time threshold δt to find the stay points. Then use the stay points to segment trajectory. Finally, SPD was verified on the geolife dataset. Mirge and Verma define the distance threshold and the angle threshold to segment trajectory [13]. Although these two algorithms can quickly find the stay points and segment the trajectory, the algorithm requires heuristic rules. In practical application, it is difficult to obtain these rules in advance, and the value of the threshold would greatly impact the accuracy of trajectory segmentation. Leiva and Vidal proposed a trajectory segmentation algorithm named Warped K-means [10] based on the K-means [29]. This algorithm adds time constraints in the K-means. It reaches 97% accuracy on real datasets. However, the number of trajectory segments k is generally unknown.

The unsupervised segmentation algorithm mainly includes clustering-based, cost-function-based, and interpolation-based. The detailed description is as follows.

The clustering-based segmentation algorithm mainly improves the existing clustering algorithm, which makes it more suitable for trajectory segmentation. A plethora of cluster-based trajectory segmentation algorithms have been proposed. CB-SMOT [27] was proposed by Andrey, which is an extension of the DBSCAN algorithm [30]. The algorithm also uses speed characteristics to discover the stop points and move points of the trajectory. And to better process the spatial-temporal trajectory data, it replaces the distance threshold in DBSCAN with the time threshold. Chen et al. improved the DBSCAN algorithm and proposed a segmentation algorithm named T-DBSCAN [18]. This algorithm utilizes the important spatial-temporal characteristics of the trajectory to segment the trajectory. The accuracy of the two algorithms is high on the experiment dataset. However, since CB-SMOT and T-DBSCAN are improved based on DBSCAN, they also have the same weaknesses as DBSCAN, which cannot reliably detect stop points from sparse trajectories.

The cost-function-based approach mainly segments the trajectory by minimizing the cost function, including GRASP-UTS [23]. It was proposed by Amilcar et al. in 2015. This algorithm first randomly selects the segmentation point, that is, landmark. Then, it utilizes the adaptive greedy algorithm to optimal the landmark and calculates the cost function. Finally, when the cost function reaches the lowest,

segment the trajectory by landmark. GRASP-UTS is tested on two real datasets of different domains and achieves high accuracy. However, because the algorithm uses an adaptive greedy algorithm, the time complexity is very high, which makes it is not suitable for large datasets.

The interpolation-based trajectory segmentation algorithm mainly uses different interpolation methods such as linear interpolation and kinematic interpolation to generate error signals for segmentation, including OWS [15] and SWS [19]. Mohammad et al. proposed the trajectory segmentation algorithm named Octal Window Segmentation (OWS) in 2019, and the SWS is an improvement of the OWS. The intuition of the two algorithms is that when a moving object changes from one behavior to another, this can be captured directly from its geographic location. Mohammad et al. compare the real position of the moving object with the estimated one to generate an error signal. By evaluating this error signal, predicting whether the behavior of moving object changed, and utilizing this information to segment trajectory. These two algorithms are better than the benchmark algorithm in segmentation accuracy. However, a part of the data is required to optimize the parameter and different trajectory datasets need to select different interpolation methods.

The semisupervised segmentation algorithm mainly includes RGRASP-SemTS [28]. RGRASP-SemTS was proposed by Amilcar et al. It uses the minimum description length (MDL) principle to measure homogeneity inside segments and segment trajectories by combining a limited user labeling phase with a low number of input parameters and no predefined segmenting criteria. However, when the algorithm faces large-scale data, it is difficult to create a part of labeled trajectory datasets.

This study proposes an efficient and accurate trajectory segmentation and merging algorithm based on multiple motion features (TS-MF) to overcome the limitations of the aftermentioned, mainly composed of a segmentation method and a trajectory merging method. The TS-MF algorithm divides the trajectory both from the global and local perspectives to ensure the accuracy of segmentation.

3. Methodology

This section details the novel unsupervised algorithm for trajectory segmentation based on multiple motion features (TS-MF). In Section 3.1, we present the relevant definitions. Figure 1 shows the overview of TS-MF, which includes the two core processing: segmentation and mergence. The first step of TS-MF is to segment the raw trajectory by Pearson coefficient, which is detailed in Section 3.2. The second step is to merge the subtrajectory of oversegmented, which is described in Section 3.3. Finally, the details of TS-MF are introduced in Section 3.4.

3.1. Definitions

3.1.1. Raw Trajectory. A raw trajectory is composed of a series of multidimensional spatial-temporal data points. It is denoted as $\text{Traj}_i = (p_0, p_1, p_2, p_3, \cdots, p_i, \cdots, p_n)\ 0 \le i \le n$, where $p_i = (\text{lat}_i, \text{lon}_i, t_i, f_i)$, lat_i and lon_i represent the position coordinates at the time t_i. f_i means the movement characteristics of the trajectory point at the time t_i such as speed, angle, and acceleration.

3.1.2. Subtrajectory. A subtrajectory is a set of consecutive trajectory points in the raw trajectory, for example, the subtrajectory can be denoted as $S_i = (p_i, p_{i+1}, \cdots, p_j)\ 0 \le i < j \le n$.

3.1.3. Trajectory Segmentation. According to feature similarity of trajectory points, the trajectory segmentation algorithm can efficiently and accurately find a set of segment points from the raw trajectory, such as $\text{Seg} = [p_0, p_1 \cdots p_k]$. We can segment the raw trajectory into several disjoint parts by these segmentation points. For example, $\text{Traj}_i = (s_0, s_1, s_2 \cdots s_k)$, where k is the number of subtrajectories.

3.2. Segmentation Method. The intuition behind the segmentation method is that when the motion features of two adjacent trajectory points (such as longitude, latitude, velocity, angle, acceleration, and heading) have significant variation, this trajectory point is where the motion state changes, that is, segmentation points. Therefore, the core of the segmentation method is to determine the segmentation point.

To accurately extract the segmentation points, it is necessary to define an index to measure the similarity of multiple motion features between two adjacent trajectory points. Since the Pearson coefficient is sensitive to variation, the Pearson coefficient is employed to calculate the similarity of adjoining trajectory points, extract the point where the motion feature changes, and save it to the segmentation point sequence.

The Pearson coefficient is a statistical indicator that reflects the degree of linear correlation between two variables. The Pearson coefficient can be calculated through Equation (1), where F_i, F_j, are the features of p_i and p_j, the features include longitude, latitude, speed, average speed, acceleration, and angle, $\text{cov}(F_i, F_j)$ is the covariance between F_i and F_j, μ represents the mean value, σ_{F_i}, σ_{F_j} means the standard deviations of F_i and F_j, and $E[(F_{i^-\mu}F_i)(F_{j^-\mu}F_j)]$ describes the expected value of $(F_{i^-\mu}F_i)(F_{j^-\mu}F_j)$. The value of ρ_{F_i,F_j} is between [-1, 1]. When ρ_{F_i,F_j} equals 0, it indicates that F_i and F_j are irrelevant; when the value equals 1 (e.g., [1–6] and [1–6]), it suggests that F_i and F_j are completely positive correlation; when the value equals -1, it means that F_i and F_j are perfectly negative correlation (e.g., [1–6] and [-1, -2, -3, -4, -5, -6]). Generally, trajectory data reflects the motion history of moving objects, and its sampling time is usually very short, so the characteristics between adjacent points in the same motion state are usually the same, that is, the value of the Pearson coefficient is close to 1. And the acceleration, speed, average speed, and angle of trajectory points with changed motion state will change obviously, resulting in Pearson coefficient is closed -1. For example, we calculate the value of the Pearson coefficient of two sets of adjacent trajectory points, whose result is shown in Table 1. We can discover that when the features

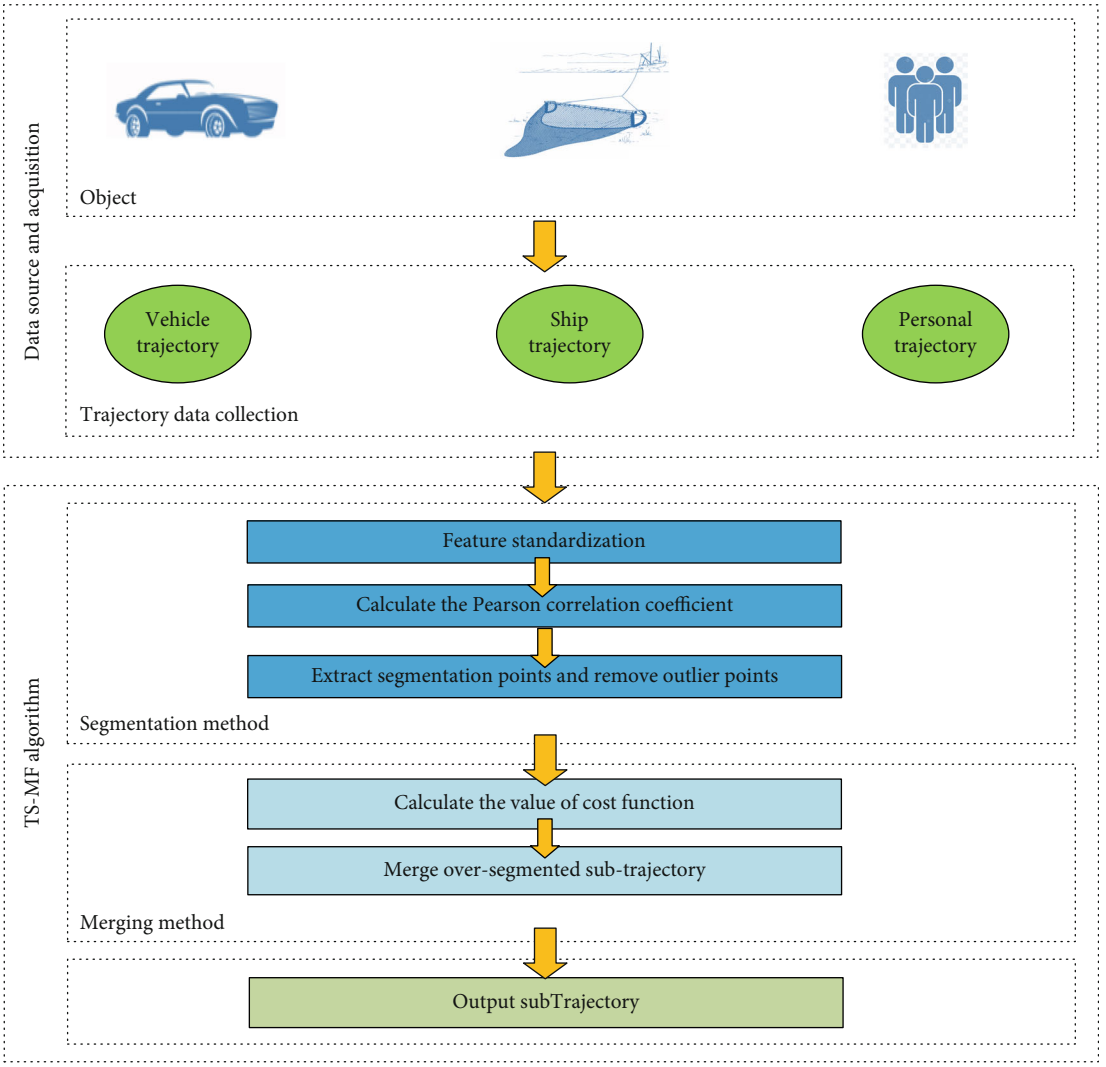

Figure 1: Overview of the proposed segmentation algorithm.

Table 1: Pearson coefficients of two sets of adjacent trajectory points.

Index	Lng	Lat	Speed	Angle	Acceleration	Average speed	ρ_{F_i,F_j}
11	-34.94	-5.59	4.39	54.7	-0.0001	4.391	0.999
12	-34.93	-5.47	4.73	38.0	0.0001	4.728	
284	-34.91	-1.284	4.43	2.66	-0.0007	4.43	−0.058
285	-34.91	-1.206	4.25	357.79	-0.0008	4.25	

of adjacent trajectory points are no obvious variation, the value of ρ_{F_i,F_j} is close to 1, and it is close to -1 otherwise.

$$\rho_{F_i,F_j} = \frac{\mathrm{cov}(F_i - F_j)}{\sigma_{F_i}\sigma_{F_j}} = \frac{E\left[\left(F_{i^{-\mu}}F_i\right)\left(F_{j^{-\mu}}F_j\right)\right]}{\sigma_{F_i}\sigma_{F_j}}. \quad (1)$$

Figure 2 shows the change of the value of ρ_{F_i,F_j}, and there are many mutation points of the Pearson coefficient. The value of ρ_{F_i,F_j} between mutation points is close to 1 and remains unchanged. Meanwhile, we can discover there are multiple mutation points in a short time. However, the motion state of the moving object does not change in a short time. It means that some multiple mutation points are the outlier points. Therefore, the purpose of the segmentation method is to extract mutation points and remove the outlier points.

The pseudocode of the segmentation method is detailed in Algorithm 1. The proposed segmentation method firstly takes out the raw trajectory (such as $\mathrm{Traj}_i = (p_0, p_1, p_2, p_3, \cdots, p_i, \cdots, p_n)\ 0 \le i \le n$) from the database. Then, calculate

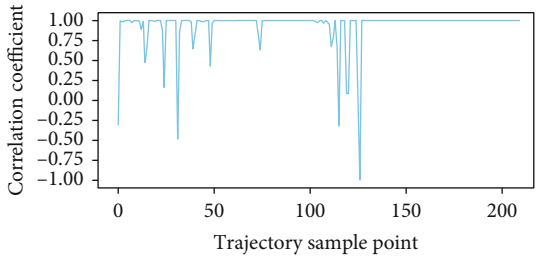

FIGURE 2: The change of the value of ρ_{F_i,F_j}.

the value of ρ_{F_i,F_j} and save the results into the array pcc. Finally, the super parameters δ and T are defined to extract the point where the motion state changes, where δ and T are the threshold of Pearson coefficient and time interval. The segmentation method looks for p_l with the minimum value of ρ_{F_i,F_j} from a global perspective. When the ρ_{F_i,F_j} less than δ and the time interval between p_l and adjacent segment points is less than T, the p_l is added the segmentation point sequence and remove the p_l from array pcc. And on the contrary, the outlier point p_l is removed. The procedure performs this step until the minimum value of ρ_{F_i,F_j} is greater than δ.

3.3. Merging Method. The trajectory segmentation algorithm based on the Pearson coefficient achieves high homogeneity in the subtrajectories. However, in practical application, the collected trajectories contain some outlier points, which cause the value of ρ_{F_i,F_j} is closed to -1. Though the segmentation method utilizes time threshold T to remove the outlier points, when the time interval is greater than T, the outlier points may be mistakenly added to the segmentation point sequence. This condition may cause the raw trajectory to be oversegmented. For example, the raw trajectory containing 122 subtrajectories is finally partitioned into 187 segments, which is oversegmented. In the mergence part, the minimum description length (MDL) principle is used to construct the cost function and merge the subtrajectories by optimizing the cost function from a local perspective, which can ensure the final segmented subtrajectory achieves the best accuracy.

The MDL was proposed by Rissanen [31] and then used and detailed by Grünwald et al. [32]. According to Grünwald et al. [32], the MDL cost consists of $L(H)$ and $L(D \mid H)$. Here, H means the hypothesis, and D the datasets. $L(H)$ is the length of the description of the hypothesis in bits, and $L(D \mid H)$ is the length of the description of the data when encoded with the hypothesis. The best hypothesis H to explain D is the one that minimizes the sum of $L(H)$ and $L(D \mid H)$.

In the problem of trajectory segmentation, a hypothesis corresponds to a subtrajectory. Finding the optimal subtrajectory means finding the best hypothesis. Give a subtrajectory $S = (p_i, p_{i+1}, \cdots, p_j)\ 0 \le i < j \le n$, and we formulate cost function by Equations (2), which can be used to measure homogeneity. In Equations (2), $L(H) = \log_2(len(p_i p_j))$ and $L(D \mid H) = \log_2(\sum_i^j d_\perp(p_i p_j, p_i p_{i+1})) + \log_2(\sum_i^j d_\theta(p_i p_j, p_i p_{i+1}))$, where $d_\perp(p_i p_j, p_i p_{i+1})$ means the perpendicular distance between $p_i p_j$ and $p_i p_{i+1}$, $d_\theta(p_i p_j, p_i p_{i+1})$ represents the angle distance between $p_i p_j$ and $p_i p_{i+1}$. The $d_\perp$ and d_θ are defined as Equations (3) and Equations (4), which are mentioned in [17]. Figure 3 shows the formulation of the cost function, $d_\perp$ and d_θ of a subtrajectory, which contains 5 trajectory points.

$$\begin{aligned} \text{cost\_function} &= L(H) + L(D \mid H) \\ &= \log_2\left(len\left(p_i p_j\right)\right) \\ &\quad + \log_2\left(\sum_i^j d_\perp\left(p_i p_j, p_i p_{i+1}\right)\right) \\ &\quad + \log_2\left(\sum_i^j d_\theta\left(p_i p_j, p_i p_{i+1}\right)\right), \end{aligned} \tag{2}$$

$$d_\perp = \frac{{l_{\perp 1}}^2 + {l_{\perp 2}}^2}{l_{\perp 1} + l_{\perp 2}}, \tag{3}$$

$$d_\theta = \|l_i\| \times \sin\theta. \tag{4}$$

Based on the theory as aforementioned, the merging method is detailed in Algorithm 2. First, the procedure uses the segmentation point sequence SegPoint to segment the raw trajectory, it can be denoted as $\text{Traj} = (s_i \cdots s_j \cdots s_k)\ 0 \le i \le j \le k$. Then mergers s_i and s_{i+1} into s_{total}, calculates the cost function of s_i, s_{i+1} and s_{total}, and the results are represented as cost_{s_i}, $\text{cost}_{s_{i+1}}$, and $\text{cost}_{s_{\text{total}}}$. When cost_{s_i}, $\text{cost}_{s_{i+1}}$, and $\text{cost}_{s_{\text{total}}}$ satisfy Equation (5), it means that the two subtrajectories are oversegmented and merge s_i and s_{i+1} from the local perspective. The procedure repeats this step until the last subtrajectory.

$$\text{cost}_{s_i} + \text{cost}_{s_{i+1}} > \text{cost}_{s_{\text{total}}}. \tag{5}$$

3.4. The TS-MF Algorithm. The segmentation part and mergence part are the two phases (global segmentation and local optimization) of TS-MF, which are described in Section 3.2 and Section 3.3. Algorithm 3 shows the pseudocode of TS-MF. This algorithm receives the following inputs: the raw trajectory $\text{Traj}_i (0 \le i \le n)$, a time threshold T and the Pearson coefficient threshold δ. The output is the set of subtrajectories, which can be denoted as $(s_i \cdots s_j \cdots s_k)$.

4. Experimental Evaluation

To evaluate the effectiveness of the proposed algorithm, we verify the proposed algorithm on two real datasets. This section first details the datasets (Section 4.1) and the evaluation metrics (Section 4.2). Then, the parameter settings and experimental results are introduced in Section 4.3 and Section 4.4, while a comparative analysis with other algorithms is presented in Section 4.5.

4.1. Trajectory Datasets. The first dataset is the vessels performing fishing activities on the coast of Brazil. It contains

Input: The raw trajectory $\text{Traj}_i (0 \le i \le n)$, time threshold T, Pearson coefficient threshold δ
Output: Segment point sequence Seg
pcc ⟵ Calculate the Pearson coefficient of all adjacent trajectory points
Seg ⟵ add p_0 and p_n to Seg
n ⟵ the sum of the num of points with $\rho_{F_i,F_j} < \delta$
for $i = 0 \longrightarrow n - 1$ **do** ∗time complexity is $\boldsymbol{O(n)}$ ∗
  ρ_{F_{l-1},F_l}, p_l ⟵ the minimum value of ρ_{F_{l-1},F_l} and the corresponding points
  p_{left}, p_{right} ⟵ the adjacent segmentation points of p_l in the SegPoint
  T_{left} ⟵ Time interval between p_l and p_{left}
  T_{right} ⟵ Time interval between p_l and p_{right}
  if $T_{left} < T$ and $T_{right} < T$ **then**
    add p_l to Seg
  end if
end for
return Seg

Algorithm 1: Trajectory segmentation algorithm based on the Pearson coefficient.

$$L(H) = \log_2(len(p_1p_5))$$
$$L(D|H) = \log_2(d_\perp(p_1p_5, p_1p_2) + d_\perp(p_1p_5, p_2p_3) + d_\perp(p_1p_5, p_3p_4) + d_\perp(p_1p_5, p_4p_5)) + \log_2(d_\theta(p_1p_5, p_1p_2) + d_\theta(p_1p_5, p_2p_3) + d_\theta(p_1p_5, p_3p_4) + d_\theta(p_1p_5, p_4p_5))$$

$$d_\perp = \frac{l_{\perp 1}^2 + l_{\perp 2}^2}{l_{\perp 1} + l_{\perp 2}}$$
$$d_\theta = ||l_i|| \cdot sin\theta$$

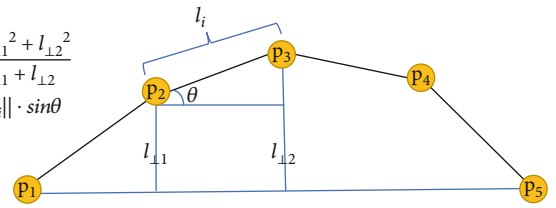

Figure 3: Schematic diagram of calculating the vertical distance, the angular distance of the line segment, and the $L(H)$ and $L(D \mid H)$ of the subtrajectory

Input: The raw trajectory $\text{Traj}_i (0 \le i \le n)$, Segment point sequence SegPoint
Output: The set of sub-trajectories $(s_i \cdots s_j \cdots s_k)$
$(s_0 \cdots s_i \cdots s_k)$ ⟵ Segment the raw trajectory Traj_i by SegPoint
k ⟵ The number of sub-trajectories
for $i = 0 \longrightarrow k$ **do**
  s_i, s_{i+1} ⟵ The s_i and its next sub-trajectory
  s_{total} ⟵ Mergers s_i and s_{i+1} into s_{total}
  $cost_{s_i}, cost_{s_{i+1}}, cost_{s_{total}}$ ⟵ Calculate the value of the loss function of s_i, s_{i+1} and s_{total}
  if $(cost_{s_i} + cost_{s_{i+1}}) > cost_{s_{total}}$ **then**
    merges s_i and s_{i+1} from the local perspective
  end if
end for
return The set of sub-trajectories $(s_i \cdots s_j \cdots s_k)$

Algorithm 2: Trajectory merging method.

5190 trajectory points and 122 segments. Our purpose is to partition the trajectories of fishing and not fishing. Generally, in Brazil, the captain must report the position (such as latitude and longitude) in real-time and record the status of fishing vessels (such as fishing and not fishing). The entire dataset was created using data from four vessels that perform the same types of fishing activities on Brazil's northeast coast. Figure 4 (left) shows the trajectory of the four fishing vessels.

The second dataset is a subset of the geolife dataset containing 12,955 trajectory points and 181 segments. The

Input: The raw trajectory $\text{Traj}_i (0 \le i \le n)$, a time threshold T, the Pearson coefficient threshold δ
Output: The set of sub-trajectories $(s_i \cdots s_j \cdots s_k)$
1: SegPoint ⟵ Segmentation method (Traj_i, T, δ)
2: $(s_i \cdots s_j \cdots s_k)$ ⟵ merging method (Traj_i, Seg)
3: **return** The set of sub-trajectories $(s_i \cdots s_j \cdots s_k)$

Algorithm 3: Trajectory segmentation based on multiple motion features (TS-MF).

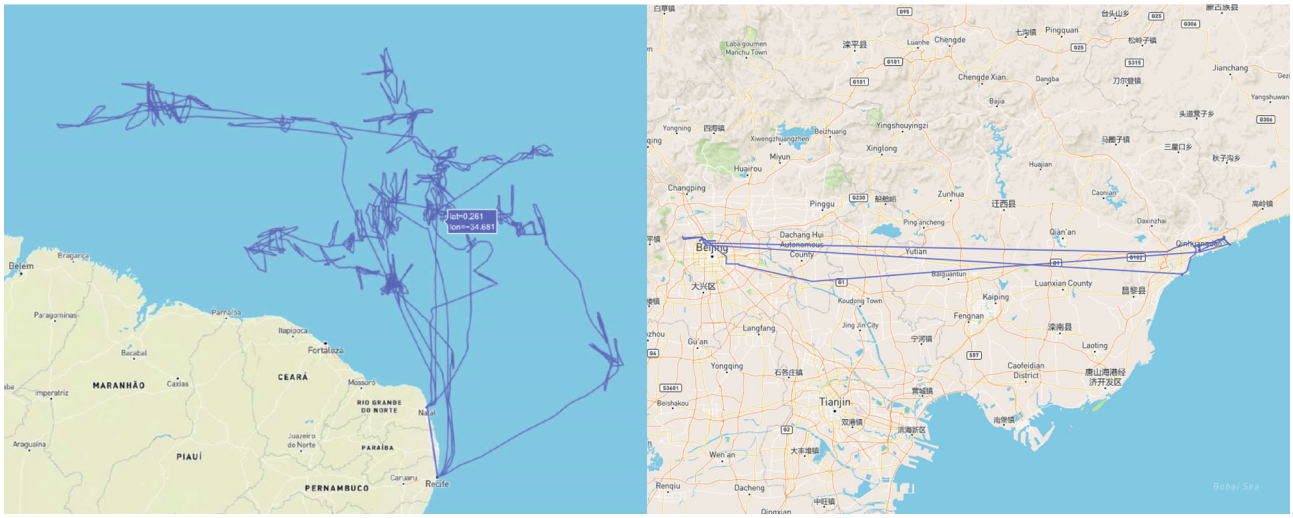

Figure 4: The fishing trajectory (left) and the part trajectory of geolife (right)

geolife dataset has a mix of behaviors, such as car, bus train, and walk. Figure 4 (right) shows the part trajectory of geolife.

From these trajectories, we extracted the information of time, longitude, latitude, fishing, speed, and angle collected. We computed some trajectory features for all the points in this dataset, including mean speed and acceleration. The data description is shown in Table 2.

4.2. Evaluation Metrics. In this work, the harmonic mean (H) of average purity $\bar{P}$ and average coverage $\bar{C}$ is used to evaluate the proposed algorithm. Scholars firstly proposed the concepts of coverage and purity in [23] and used the harmonic mean (H) to evaluate the trajectory segmentation algorithm in [19].

The segment purity is the ratio of the sum of the most frequent label in the segment and the sum of all the trajectory points. For example, suppose a segmented trajectory has k points, and the number of trajectory points with the most same label is d, then, the segment purity C is d/k. The average of purity values for all segments is called as $\bar{P}$. Coverage is to evaluate the completeness of the segmentation algorithm. For example, suppose that the raw trajectory segment τ is divided into τ_1, τ_2, τ_2 is the larger one, and the coverage C is defined as τ_2/τ. The average for coverage of all segments is called as $\bar{C}$. Since the two metrics of purity (P) and coverage (C) are designed to be orthogonal, i.e., when one index increases, the other index decreases. Therefore,

Table 2: Data description.

Field name	Field description	Sample data
Lng	Longitude	-34.9
Lat	Latitude	0.861
Time	Timestamp	1082199600
Label	Operation status	Fishing
Speed	Speed	4.5
Angle	Angle	358.3155
Accel	Accelerated	0.000139
Avg	Mean speed	4.502692

the harmonic mean of the purity and coverage is used to evaluate the performance of TS-MF. Equation (6) gives the formulation of the harmonic mean [19]. When the harmonic mean is the highest, the purity and coverage of the segmented trajectory reach a good compromise, and the segmentation of subtrajectories is the best.

$$H = \frac{2 * \bar{P} * \bar{C}}{\bar{P} + \bar{C}}. \tag{6}$$

4.3. Parameter Settings. In the segmentation process, the threshold of Pearson coefficient, that is, σ is employed to find the segmentation point. In general, when $\rho_{F_i,F_j} \ge 0.8$, the two variables of features are highly positively correlated,

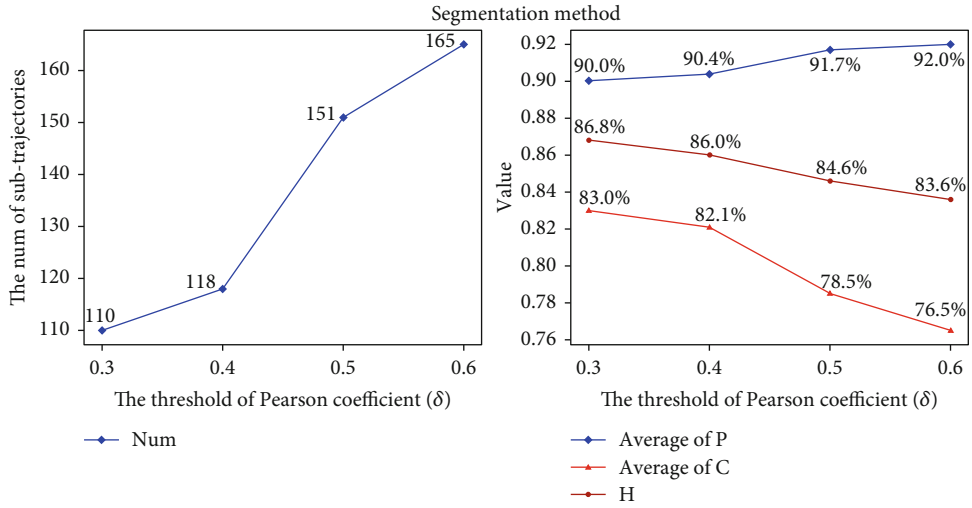

Figure 5: The evaluation results of the segmentation method on fishing dataset

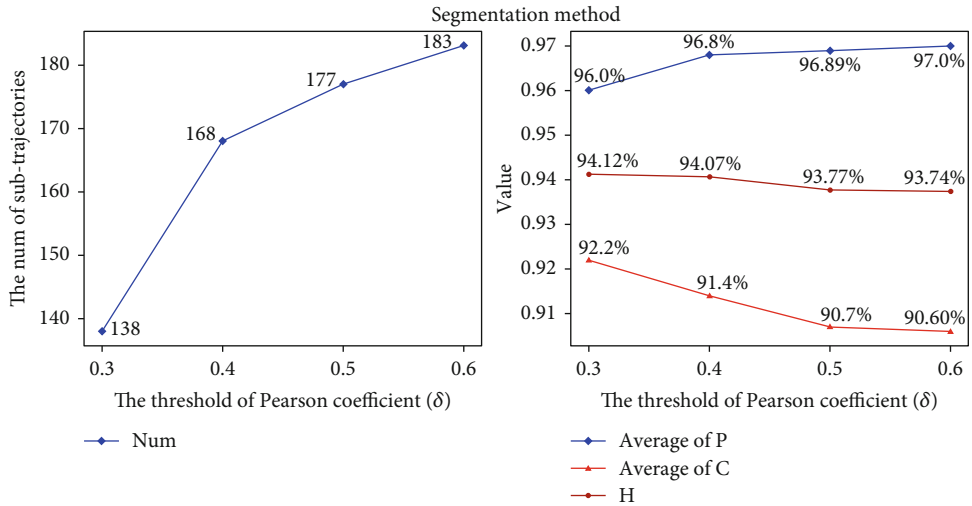

Figure 6: The evaluation results of the segmentation method on geolife dataset

$0.6 \le \rho_{F_i,F_j} < 0.8$, the two variables of features are moderately positive correlated; $0.4 \le \rho_{F_i,F_j} < 0.6$, the two variables of features can be low correlation; $0 \le \rho_{F_i,F_j} < 0.4$, the two variables of features may be irrelevant, and $\rho_{F_i,F_j} < 0$ suggests that two variables of features are negatively correlated. Therefore, the TS-MF can make $\delta = 0.4$ to extract segmentation points. In addition, the segmentation process also utilizes the threshold of T to remove outlier points. Since it is difficult to know the specific duration of each state of the moving object and the purpose of setting the T is only to remove the part outlier points, the T can be set to the minimum value of the duration of each movement state. The duration of fishing activities of fishing vessels is 6 hours on the coast of Brazil, which is mentioned in [33], and the shortest duration of the walk generally is 30 min. Therefore, the $T = 6$ hours for the vessels performing fishing activities on the coast of Brazil and $T = 0.5$ hours for the geolife.

4.4. Experiment Result and Analysis. The experiment result and analysis are detailed in this section. In this experiment, the segmentation method and TS-MF are evaluated on the fishing dataset and geolife dataset. In addition, to observe the impacts of δ and demonstrate the feasibility of $\delta = 0.4$, it is tested under different δ. Figures 5–8 show the result of the experiment.

The results are shown in Figures 5–8, which display the value of the sum of subtrajectories, $\bar{P}$, $\bar{C}$, and H under different σ on different datasets.

The results of the segmentation method are shown in Figures 5–6. The results display that with the increase of δ, the sum of subtrajectories increases, the $\bar{P}$ increases, the $\bar{C}$ decreases, and the H increases.

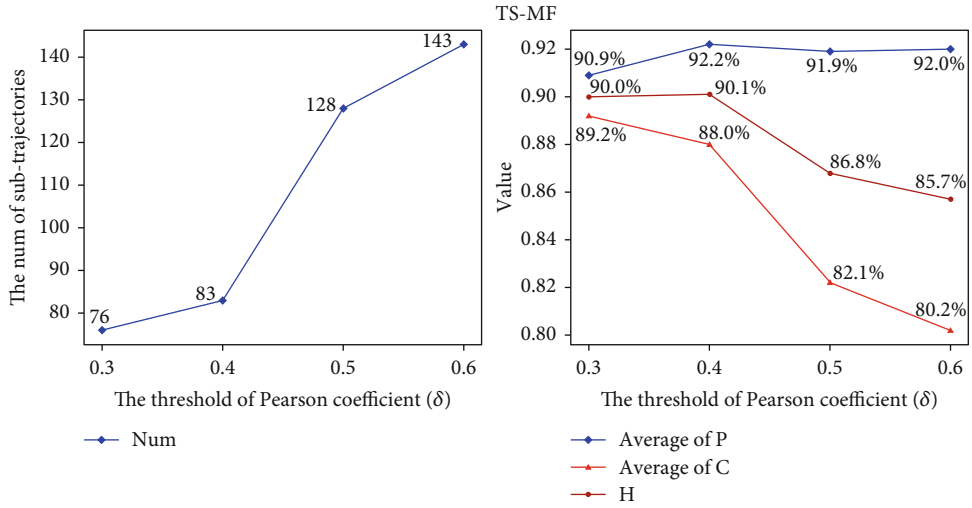

Figure 7: The evaluation results of TS-MF on fishing dataset

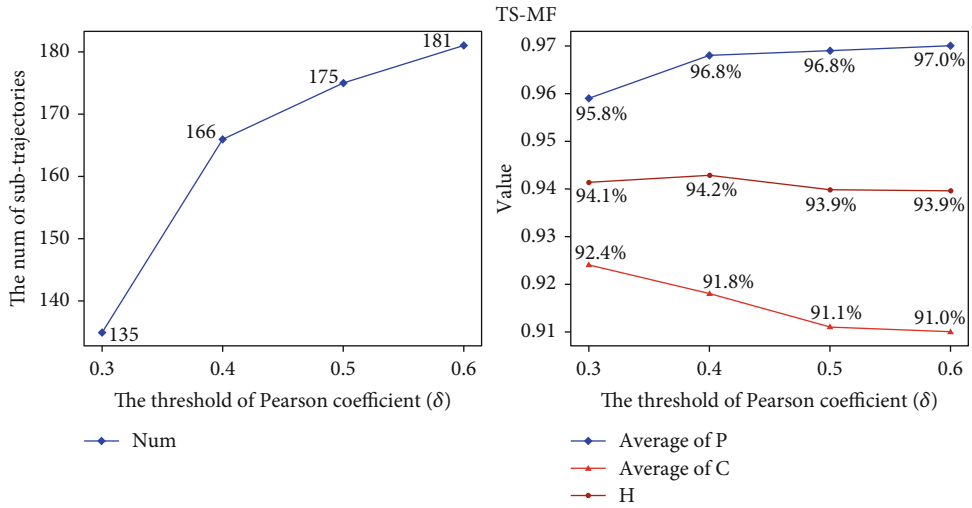

Figure 8: The evaluation results of TS-MF on geolife dataset

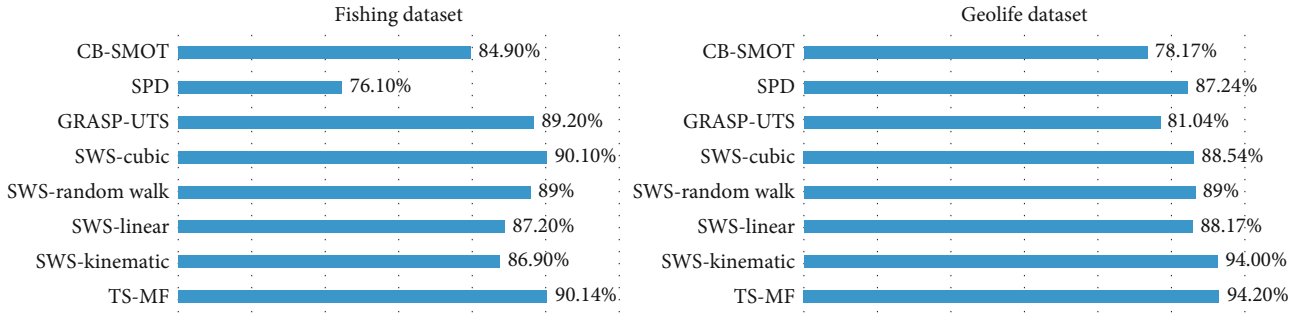

Figure 9: Compare with other segmentation algorithms on fishing dataset (left) and geolife dataset (right)

TS-MF is an extension of the segmentation method, that is, there is one more merging method. The mergence part merges the local subtrajectory by the segmentation method. The results of TS-MF are as shown in Figures 7–8. Compare the results of the segmentation method, we can observe the num of subtrajectory is lower and the $\bar{C}$ and H is better. We also can discover that in the mergence part, the num of merged subtrajectories on the fishing dataset is more than

geolife dataset. The reason is that when fishing vessels engage in fishing, the speed generally is 4 miles per hour, and the heading angle is constantly changing. This condition leads to the value of the Pearson coefficient being lower and the segmentation method may add many outlier points into segmentation points. Geolife collected trajectory data of 182 users, which includes various motion states. The difference of features of different motion states is large while the same motion state is small. Therefore, the segmentation method can accurately discover the segmentation points, that is, the outlier points in segmentation points is less.

Overall, the results of TS-MF are better and the greater of δ, the segmentation method can extract more segmentation points and leads to the $\bar{C}$ and H becomes lower. But it does not mean that the lowest δ is the best selection. As shown in Figures 7–8, when the $\delta = 0.3$, the sum of subtrajectory is very low, that is, many segmentation points are lost of TS-MF. The results also indicate that it is the feasibility of $\delta = 0.4$.

4.5. Comparing TS-MF with Other Baseline Algorithms. In this section, the experiment is repeated in the same environment, and TS-MF was compared with the other four trajectory segmentation algorithms (CB-SMoT, GRASP-UTS, SPD, and SWS) on the fishing dataset and geolife dataset. The results are reported in Figure 9. As shown in Figure 9, we can discover that the value of harmonic average is 90.1% and 94.28% on different datasets, and TS-MF achieves the highest harmonic average of purity and coverage. The results also demonstrate the feasibility of TS-MF.

5. Conclusions

It is envisioned that future wireless communications will be more data-driven. It is possible to obtain the high-accuracy and long-term trajectory of a moving object by mobile edge cloud, beamforming, and artificial intelligence techniques. But the long-term location data need huge computing resources to process and loses a lot of information. The segmentation algorithm designed for location data is the basic step to develop the location-based application. This study proposes an unsupervised trajectory segmentation algorithm, named TS-MF, which employs the Pearson coefficient to find the segmentation points and minimum cost function to merge the oversegmented subtrajectory. We compared our proposed segmentation algorithm against GRASP-UTS, SPD, CB-SMoT, and SWS; the results show that the proposed algorithm reaches the best harmonic mean of purity and coverage on the fishing dataset and geolife dataset. Furthermore, the TS-MF algorithm requires no labeled data and its time complexity is $O(n)$, which means it is computation efficient and thus most suitable for the segmentation of large trajectory datasets.

However, there is one limitation of TS-MF. It is that when the features are similar in the different movement states, the proposed segmentation algorithm may not find the qualified segmentation points for the raw trajectory.

As future work, we plan to extend this work in other directions. First, we would analyze the trajectory motion pattern and predict the subtrajectory state, semantic enhancement for raw trajectory. Second, we would like to apply the segmentation algorithm (TS-MF) to more wireless positioning data, which facilitates more artificial intelligence technology are used to mine valuable information.

Data Availability

The data and codes that support the findings of this study are available with the identifier(s) at the private link https://figshare.com/s/6e6fb483b076b2a34cbe.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Wenjin Xu and Shaokang Dong conceived and designed the experiments; Shaokang Dong performed the experiments and analyzed the data; Wenjin Xu and Shaokang Dong wrote the paper. Wenjin Xu and Shaokang Dong contributed equally to this work.

References

[1] M. Etemad, A. Júnior, and S. Matwin, "Predicting transportation modes of GPS trajectories using feature engineering and noise removal," in *Canadian Conference on Artificial Intelligence*, Lecture Notes in Computer Science, Springer, Cham, 2018.

[2] E. N. de Souza, K. Boerder, S. Matwin, and B. Worm, "Correction: improving fishing pattern detection from satellite AIS using data mining and machine learning," *PLoS One*, vol. 11, no. 9, 2016.

[3] J. Zhang and X. Zhao, "Distributed power allocation for cognitive radio networks with time varying channel and delay: H_∞ state feedback control approach," *IEEE Access*, vol. 6, pp. 56893–56910, 2018.

[4] C. Calenge, S. Dray, and M. Royer-Carenzi, "The concept of animals' trajectories from a data analysis perspective," *Ecological Informatics*, vol. 4, no. 1, pp. 34–41, 2009.

[5] P. Bovet and S. Benhamou, "Spatial analysis of animals' movements using a correlated random walk model," *Journal of Theoretical Biology*, vol. 131, no. 4, pp. 419–433, 1988.

[6] E. Gurarie, R. D. Andrews, and K. L. Laidre, "A novel method for identifying behavioural changes in animal movement data," *Ecology Letters*, vol. 12, no. 5, pp. 395–408, 2009.

[7] M. Yue, Y. Li, H. Yang, R. Ahuja, Y. Y. Chiang, and C. Shahabi, "DETECT: deep trajectory clustering for mobility-behavior analysis," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 988–997, Los Angeles, CA, USA, 2020.

[8] Y. Zheng, "Trajectory data mining," *ACM Transactions on Intelligent Systems and Technology*, vol. 6, no. 3, pp. 1–41, 2015.

[9] M. Etemad, Z. Etemad, A. Soares, V. Bogorny, S. Matwin, and L. Torgo, "Wise sliding window segmentation: a classification-aided approach for trajectory segmentation," in *Canadian Conference on Artificial Intelligence*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020.

[10] L. A. Leiva and E. Vidal, "Warped *K*-means: an algorithm to cluster sequentially-distributed data," *Information Sciences*, vol. 237, pp. 196–210, 2013.

[11] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Y. Ma, "Recommending friends and locations based on individual location history," *ACM Transactions on the Web (TWEB)*, vol. 5, no. 1, pp. 1–44, 2011.

[12] A. Anagnostopoulos, M. Vlachos, M. Hadjieleftheriou, E. Keogh, and P. S. Yu, "Global distance-based segmentation of trajectories," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 34–43, Philadelphia, PA, USA, 2006.

[13] V. Mirge and K. Verma, "Distance and bearing based vehicle trajectory segmentation," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 4, pp. 677–681, 2019.

[14] M. Buchin, A. Driemel, M. Van Kreveld, and V. Sacristán, "An algorithmic framework for segmenting trajectories based on spatio-temporal criteria," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 202–211, San Jose, California, 2010.

[15] M. Etemad, A. S. Júnior, A. Hoseyni, J. Rose, and S. Matwin, "A trajectory segmentation algorithm based on interpolation-based change detection strategies," in *EDBT/ICDT Workshops*, International Conference on Extending Database Technology/International Conference on Database Theory, 2019.

[16] J. Tang, L. Liu, and J. Wu, "A trajectory partition method based on combined movement features," *Wireless Communications and Mobile Computing*, vol. 2019, 7803293:13 pages, 2019.

[17] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 593–604, Beijing, China, 2007.

[18] W. Chen, M. Ji, and J. Wang, "T-DBSCAN: a spatiotemporal density clustering for GPS trajectory segmentation," *International Journal of Online Engineering*, vol. 10, no. 6, pp. 19–24, 2014.

[19] M. Etemad, A. Soares, E. Etemad, J. Rose, L. Torgo, and S. Matwin, "SWS: an unsupervised trajectory segmentation algorithm based on change detection with interpolation kernels," *GeoInformatica*, vol. 25, pp. 269–289, 2021.

[20] Y. Gao, L. Huang, J. Feng, and X. Wang, "Semantic trajectory segmentation based on change-point detection and ontology," *International Journal of Geographical Information Science*, vol. 34, no. 12, pp. 2361–2394, 2020.

[21] A. Bonavita, R. Guidotti, and M. Nanni, "Self-adapting trajectory segmentation," in *EDBT/ICDT Workshops*, International Conference on Extending Database Technology/International Conference on Database Theory, 2020.

[22] S. Hwang, C. VanDeMark, N. Dhatt, S. V. Yalla, and R. T. Crews, "Segmenting human trajectory data by movement states while addressing signal loss and signal noise," *International Journal of Geographical Information Science*, vol. 32, no. 7, pp. 1391–1412, 2018.

[23] A. Soares Júnior, B. N. Moreno, V. C. Times, S. Matwin, and L. A. F. Cabral, "GRASP-UTS: an algorithm for unsupervised trajectory segmentation," *International Journal of Geographical Information Science*, vol. 29, no. 1, pp. 46–68, 2015.

[24] S. Guo, X. Li, W. K. Ching, R. Dan, W. K. Li, and Z. Zhang, "GPS trajectory data segmentation based on probabilistic logic," *International Journal of Approximate Reasoning*, vol. 103, pp. 227–247, 2018.

[25] Z. Izakian, M. S. Mesgari, and R. Weibel, "A feature extraction based trajectory segmentation approach based on multiple movement parameters," *Engineering Applications of Artificial Intelligence*, vol. 88, article 103394, 2020.

[26] S. Moosavi, R. Ramnath, and A. Nandi, "Discovery of driving patterns by trajectory segmentation," in *Proceedings of the 3rd ACM SIGSPATIAL PhD Symposium*, pp. 1–4, Burlingame, California, 2016.

[27] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares, "A clustering-based approach for discovering interesting places in trajectories," in *Proceedings of the 2008 ACM symposium on Applied computing*, pp. 863–868, Fortaleza, Ceara, Brazil, 2008.

[28] A. S. Júnior, V. C. Times, C. Renso, S. Matwin, and L. A. Cabral, "A semi-supervised approach for the semantic segmentation of trajectories," in *2018 19th IEEE International Conference on Mobile Data Management (MDM)*, pp. 145–154, Aalborg, Denmark, 2018.

[29] J. B. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, California, 1967.

[30] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *kdd*, vol. 96, no. 34, pp. 226–231, 1996.

[31] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[32] P. Grünwald, I. J. Myung, and M. Pitt, *Advances in Minimum Description Length: Theory and Applications*, MIT Press, Cambridge, MA, 2005.

[33] J. A. M. Rocha, V. C. Times, G. Oliveira, L. O. Alvares, and V. Bogorny, "DB-SMoT: a direction-based spatio-temporal clustering method," in *2010 5th IEEE international conference intelligent systems*, pp. 114–119, London, UK, 2010.