International Journal of Geo-Information

MDPI

Article

Detecting Anomalous Trajectories and Behavior Patterns Using Hierarchical Clustering from Taxi GPS Data

Yulong Wang [1], Kun Qin [1,2,*], Yixiang Chen [3] and Pengxiang Zhao [4]

1 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; wangyulong@whu.edu.cn
2 Collaborative Innovation Center for Geospatial Technology, Wuhan University, Wuhan 430079, China
3 Department of Surveying and Geoinformatics, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; chenyixiang@njupt.edu.cn
4 Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon 999077, Hong Kong, China; peng.x.zhao@polyu.edu.hk
* Correspondence: qink@whu.edu.cn; Tel.: +86-027-6877-8392

Received: 3 November 2017; Accepted: 11 January 2018; Published: 12 January 2018

Abstract: Anomalous taxi trajectories are those chosen by a small number of drivers that are different from the regular choices of other drivers. These anomalous driving trajectories provide us an opportunity to extract driver or passenger behaviors and monitor adverse urban traffic events. Because various trajectory clustering methods have previously proven to be an effective means to analyze similarities and anomalies within taxi GPS trajectory data, we focus on the problem of detecting anomalous taxi trajectories, and we develop our trajectory clustering method based on the edit distance and hierarchical clustering. To achieve this objective, first, we obtain all the taxi trajectories crossing the same source–destination pairs from taxi trajectories and take these trajectories as clustering objects. Second, an edit distance algorithm is modified to measure the similarity of the trajectories. Then, we distinguish regular trajectories and anomalous trajectories by applying adaptive hierarchical clustering based on an optimal number of clusters. Moreover, we further analyze these anomalous trajectories and discover four anomalous behavior patterns to speculate on the cause of an anomaly based on statistical indicators of time and length. The experimental results show that the proposed method can effectively detect anomalous trajectories and can be used to infer clearly fraudulent driving routes and the occurrence of adverse traffic events.

Keywords: trajectory clustering; trajectory anomalies; edit distance; hierarchical clustering; anomalous behavior pattern

1. Introduction

As a part of urban public transport, taxi service has played a positive role in promoting urban economic development and convenience in our daily lives [1]. Meanwhile, it is associated with problems such as traffic congestion, taxi fraud and detours, refusal to take passengers, etc. Especially in China, the irrational taxi phenomenon is particularly prominent in those cities with a large area and population and a sophisticated road network. A traditional means of addressing this problem would require regular random sampling, which would require significant human resources and economic costs. At present, an increasing number of vehicles are equipped with GPS navigation equipment [2]. Thousands of taxis periodically report their positions, directions, and speed as pervasive sensors of the road network each day, thereby creating a massive amount of trajectory data over time [3], which contain interesting and unexpected information about urban traffic systems [4]. Fortunately, based on this information, we can hopefully detect the aggregation or isolation trajectories in time and space [5].

Most previous studies have proposed anomaly detection using trajectory data, which can be analyzed in two ways. One approach is the study of anomalous traffic. Relevant research works regard trajectories as a traffic flow and consider anomalous traffic to be those areas or roads where the values of corresponding traffic indicators deviate from the expected value [6–10]. Through analysis of anomalous traffic, anomalous events occurring in urban traffic can be detected, and large-scale traffic conditions can be monitored [6–9]. This analysis also provides the possibility to explore the root cause of an anomalous trajectory [10]. The second approach is the study of anomalous trajectories. Relevant researchers study the attribute information of trajectory data and aim to distinguish between the minority choices and majority choices of drivers. The anomalous trajectories are the minority choices of drivers. Several data mining methods have been proposed to achieve the goals of monitoring the behavior of taxi drivers, especially fraudulent driver behavior, and recommending dynamic routes based on road conditions or destinations [5,10–12].

Our research focuses on anomalous trajectory detection. The data used in these relevant works include GPS data, social data, video data, etc. [13–16]. The relevant methods can be classified as the statistical method, distance-based method, clustering-based method, or classification-based method [12,17]. This research uses the clustering-based method and taxi GPS trajectory data. Two challenges remain in anomalous trajectory detection research [18]. The first challenge is calculating the similarity of trajectories with a true GPS location. Because trajectory data are time series data, many similar methods for time series can be improved and adopted. In addition, according to the different purposes of the application, a trajectory can be divided into sub-trajectories, or one can express a trajectory in another form to replace a coordinate pair calculation, which includes a dividing grid or expression of the road network. The second challenge is selecting a suitable clustering algorithm without prior knowledge to make the method appropriate for trajectory data. The clustering method unavoidably uses parameters have been selected based on experience or multiple attempts. Moreover, different clustering numbers will produce different results. The method of clustering number determination needs to be improved based on the features of the line segment.

In this paper, a trajectory clustering method based on edit distance and hierarchical clustering is proposed to detect anomalous trajectories. The editing distance algorithm can be used to calculate the similarity of trajectory data. It is necessary to identify the operating cost of the edit distance based on the characteristics of GPS trajectories. The hierarchical clustering method can cluster trajectories into groups, and it is necessary to determine the clustering number. Sum-of-squares-based indices show promising properties in terms of determining the number of clusters and can be improved to be suitable for the evaluation of trajectory clusters. Experimental data collected from Wuhan city taxis are used to detect anomalous trajectories. The dataset provides an extensive amount of taxi trajectory data, recording the taxi number, time, velocity, geo-location, and other attribute information. The results show that the proposed method can effectively detect anomalous trajectories. In addition, we further analyze the anomalous trajectories and determine four anomalous behavior patterns. These anomalous patterns summarize the reasons for the anomalous trajectories, which are highly significant in taxi driver supervision and traffic management.

The article is organized as follows: we review related work regarding trajectory clustering for anomalous trajectory detection in Section 2. Then, Section 3 describes the proposed method for detecting anomalous trajectories. Section 4 presents a series of experiments on anomalous trajectory detection and anomalous trajectory behavior pattern analysis, demonstrating the advantages and effectiveness of the proposed approach. We discuss the research results in Section 5 and conclude this work in Section 6.

2. Related Works about Trajectory Clustering for Anomalous Trajectory Detection

As previously mentioned, the trajectory clustering method can be used to detect anomalous trajectories. According to the clustering target, existing approaches could be divided into whole trajectory clustering and sub-trajectory clustering. Previous works divide trajectory data into

sub-trajectory datasets and take these sub-trajectories as the clustering object. Lee et al. [19] first partitioned a trajectory into a set of line segments, then calculated the distances of the line segments based on the three components of Euclidean distance. Final detection of anomalous sub-trajectories was done using a hybrid of the distance-based and density-based approaches. Guan et al. [20] defined the distance between sub-trajectories based on local and relative distance from the line Hausdorff distance. In addition, an R-tree was employed to improve the efficiency of the DBSCAN method. These distance algorithms are effective for a similarity measurement of a sub trajectory. However, they ignored the integrality of trajectory behavior information and may not be applicable to the whole trajectory.

There has been another method to express trajectories in another form to replace coordinate pair calculation. Won et al. [21] proposed a new clustering scheme for objects moving on road networks. First, trajectory data can be represented as a sequence of road segments. Then, a similarity measurement of a trajectory segment based on DSN (dissimilarity with number) is defined. Finally, one chooses hierarchical clustering as a trajectory query-processing scheme. Sha et al. [22] focused on spatiotemporally similar trajectories of road networks and defined the similarity between the query locations and the trajectory on the road networks based on a Network Voronoi Diagram. Zhang et al. [12] first divided the city into grid cells of equal size, then taxi trajectories are represented in cell grid and become a sequence of traversed cells. Secondly, isolation forest method is developed and a data-induced random tree (iTree) is utilized. A randomly selected cell is used to recursively divide the data in each node of the iTree until the node has only one trajectory or all trajectories at the node are the same. Finally, the trajectories that have short path lengths in iTree are suspected to be anomalous. The trajectory expression of the road network is related to the structure of the road network. When the density of a road network is low, some true locations of trajectories will be lost. Otherwise, the grid size affects the accuracy of the trajectory expression and the computational cost of the algorithm.

Taking a whole trajectory as a clustering object, a key step is defining the similarity measurement between different trajectories based on the characteristics of the trajectory data. There have been many methods used to realize a similarity measurement, including Euclidean distance [23], Hausdorff distance [20], LCSS (longest common subsequences) [24], DTW (dynamic time warping) [7], ERP (edit distance with real penalty) [25], and EDR (edit distance in real sequence) [26]. In practical application, the Euclidean distance requires two trajectories of the same length. Hausdorff distance does not consider the structural relationship between trajectories. LCSS requires reasonable similar threshold parameter selection. DTW is sensitive to noise and cannot effectively identify the dissimilarity of small part interval dissimilarity. However, edit distance, which is the similarity measurement method of multi sub-time interval correspondence, does not require correspondence between points and the points of the two trajectories, which can reflect the structural differences between the trajectory sequences and determine the similarities of whole trajectories. At present, there are few studies on the application of edit distance to GPS trajectories [27,28]. The specific situation of edit distance will be discussed in Section 3.1.

The selection of a suitable clustering algorithm includes a density-based method [20,23,29], spectral clustering [30], a model-based method [12], or hierarchical clustering [31]. Lee et al. [23] chose a density-based line-segment clustering algorithm for grouping similar line segments together. Bermingham et al. [29] improved TRACLUS to create ND-TRACLUS, which introduced Retraspam, and split and merged the most representative line segments simply. Guan et al. [20] proposed trajectory clustering based on an improved minimum Hausdorff distance, named TraClustMHD. Fu et al. [30] first resampled trajectories by equal space intervals. Then, they employed spectral clustering to group trajectories of similar spatial patterns. Roh et al. [31] proposed a new trajectory clustering called NNCluster, a modified agglomerative hierarchical clustering that was chosen as a baseline algorithm, to reduce the number of distance computations during the clustering process. Density-based clustering methods need to calculate the density between objects. For trajectory data, the density calculation

needs to regard trajectories as a point object, which will result in a loss of trajectory integrity, and it will not be easy to identify the trajectory represented by the points. Spectral clustering is not good for processing high-dimensional data and depends on a similarity matrix. Different similarity matrices may result in different clustering results.

In summary, existing anomalous trajectory detection methods based on trajectory clustering should either calculate the similarity of trajectories with many parameters or determine the number of clusters that requires prior knowledge participation. The current study develops a new trajectory clustering method and applies it to taxi trajectory data to detect anomalous trajectories.

3. Trajectory Clustering Method that Integrates Edit Distance and Hierarchical Clustering

A procedure for anomalous trajectory detection is introduced in Figure 1. There have given a list of N trajectories $T = \{T_1, T_2, \ldots, T_N\}$. There are three main aspects to consider in the procedure. The first aspect is the similarity of trajectories based on edit distance to obtain the distance matrix, which can be employed to obtain the clusters by using a hierarchical clustering method. The second aspect is how to determine the cluster number of the hierarchical clustering. These clusters include normal trajectories and anomalous trajectories. The third aspect is to further divide detected anomalous trajectories into different anomalous behavior patterns based on a statistical indicator.

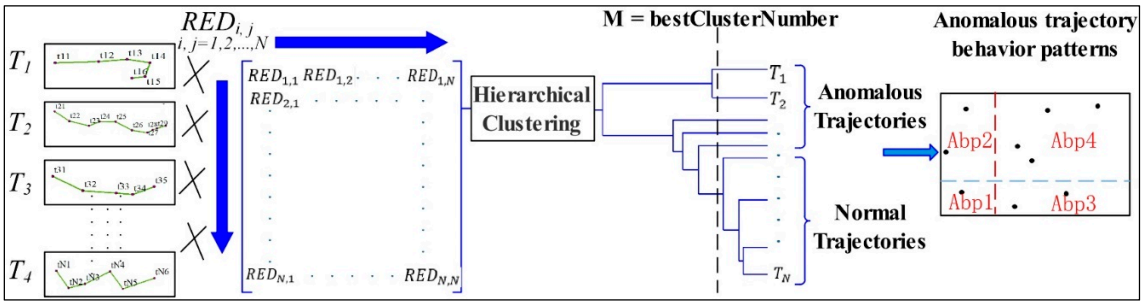


Figure 1. A procedure for the anomalous trajectory detection.

3.1. Improved Edit Distance for GPS Trajectory Data

The edit distance algorithm has proven success in assessing trajectory similarity in many research projects. Chen et al. [25,26] earlier applied edit distance to time series data and proposed ERP (edit distance with real penalty) and EDR (edit distance on real sequence) functions to measure the similarity between two trajectories. The matching thresholds were defined to reduce the effects of noise, and subcost was defined to address local time shifting. However, these methods take into account the general movement trajectory data, and the operation cost is only 1. Because the sampling time of the GPS data is inconsistent, the distance between the sampling points is different, and the operation cost must be modified. Dodge et al. [27] represented trajectories by a string representation according to movement parameters, such as speed, acceleration, or direction. However, the method focused on the parameters describing the dynamic characteristics of movement and did not address trajectory similarity for geospatial space. Yuan et al. [28] modified the operation cost based on the effects of each operation by measuring the centroid displacement after each operation from detailed call record data. The centroid of the trajectory by calculating the average position. However, taxi trajectory data are restricted by road networks, and the centroid of taxi trajectory may not be on the road. The centroid method cannot be used directly and needs to be improved.

In our study, an improved real edit distance operation cost is proposed, which include two aspects: (1) An edit distance value of a taxi trajectory must identify the point operation that corresponds to the coordinates. Thus, we need to redefine the operation cost of the corresponding coordinates. Considering that trajectory data are recorded in a time series, the point of current operation has a significant relationship with the point of previous record. Thus, the operation cost is defined based

on the gap between the position point of current operation and the previous position point. The value of the operation cost is calculated based on real coordinate positions between two operation points. (2) The time and length of each trajectory record are different. Because the edit distance value is related to the number of record points contained in the trajectory, this leads to the edit distance value of a long-sequence trajectory being greater than the edit distance value of a short-sequence trajectory. To solve this problem, the edit distance value needs to be normalized.

3.1.1. Edit Distance Operation Cost

Given two real trajectory sequences of moving objects $R(r_1, r_2, \cdots, r_m)$ and $S(s_1, s_2, \cdots, s_n)$, the defined formula $IED(R, S)$ of the revised edit distance to transform the R sequence to the S sequence is

$$\min \begin{pmatrix} \sum_{j=1}^{n} Cost[insert(s_j)] & m = 0 \\ \sum_{i=1}^{m} Cost[delete(r_i)] & n = 0 \\ IED(Rest(R), S) + Cost[delete(r_m)], & \\ IED(R, Rest(S)) + Cost[insert(s_n)], & other \\ IED(Rest(R), Rest(S)) + Cost[replace(r_m, s_n)] & \end{pmatrix} . \quad (1)$$

Equation (1) is a recursive formula; hence, m is a length of the R sequence, n is a length of the S sequence, r_i is the i element of the R sequence, and s_j is the j element of the S sequence. $Rest(R) = \{r_1, r_2, \cdots, r_{m-1}\}$ are the other parts of the R sequence removing the current point, and $Rest(S) = \{s_1, s_2, \cdots, s_{n-1}\}$ are the other parts of the S sequence removing the current point.

The operations include insert, delete and replace. For trajectories $R(r_1, r_2, \cdots, r_m)$ and $S(s_1, s_2, \cdots, s_n)$, r_i and s_j contain the actual coordinate positions (x_i, y_i) and (x_j, y_j). Every cost function is defined as

$$Cost[insert(s_j)] = |s_j - s_{j-1}| \qquad j > 1 \quad (2)$$

$$Cost[delete(r_i)] = \begin{cases} |r_i - r_{i-1}| & i > 1 \quad m = 0 \\ |r_i - s_j| & m \neq 0 \end{cases} , \quad (3)$$

$$Cost[replace(r_m, s_n)] = \begin{cases} |r_i - s_j| & |r_i - s_j| > \theta \\ 0 & |r_i - s_j| < \theta \end{cases} . \quad (4)$$

The process aims to transform an R sequence to an S sequence. Thus, all operations are taken on the original R sequence. Operation cost can be calculated based on Equations (2)–(4). However, if we transform an S sequence to an R sequence, the values of operation cost differ from the values when we transform an R sequence to an S sequence. Hence, the edit distance values of $IED(R, S)$ and $IED(S, R)$ are different. The asymmetry of this distance is reasonable in the field of cognitive science. However, in our study, we focus on the physical meaning of edit distance using the edit distance algorithm to measure the similarity between two trajectories. Thus, the edit distance between trajectories R and S does not consider asymmetry and can calculate the average value of $IED(R, S)$ and $IED(S, R)$.

3.1.2. Normalization

The formula for edit distance normalization is

$$NIED(R, S) = \frac{IED(R, S)}{IED\_max(R) + IED\_max(S)} \quad (5)$$

In formula (5), $IED(R, S)$ is the edit distance of trajectories R and S. $IED\_\max(R)$ is the length of the continuous points of trajectory R, and $IED\_\max(S)$ is the length of the continuous points of trajectory S. The denominator of the formula, which denotes that the two trajectories are completely irrelevant, is the maximum distance weight required for the operation from one trajectory to another.

When edit distance is normalized between 0 and 1, 0 denotes that the two trajectories are exactly the same, whereas 1 indicates that the two trajectories are completely irrelevant.

3.2. Determining the Number of Clusters in Hierarchical Clustering

In hierarchical clustering, clusters are merged or divided according to linkage standards, which include single linkage, complete linkage, average linkage, centroid method, and Ward's method [32]. Different linkage standards can lead to various clustering results. When the distance matrix is calculated, single linkage, complete linkage, and average linkage are relatively easy to calculate. The trajectory is composed of a series of points, and the centroid of the trajectory can be calculated as the average location of these points [28]. Ward's method calculates the sum of square errors generated by the merging of two clusters. It is necessary to determine which method to adopt by experiments.

Determining the number of clusters is an important part of hierarchical cluster analysis. The classic method is the elbow point method, which is described by Ketchen [33] and presented in Yuan and Raubal [28]. The number of clusters and the merge distance are fitted to a curve. In general, the elbow point appears at the point of maximum value change in the tangent slope. However, in real-world applications, the 'elbow point' cannot always be unambiguously identified [33]. Other methods, such as the L-method [34], have been proposed to identify the elbow point of the curve by examining the boundary between the pair of straight lines that most closely fit the curve. Zhao et al. [35] considered sum-of-squares-based indices that show promising properties in terms of determining the number of clusters. Therefore, the WB-index was proposed and had a minimum value as the determined number of clusters. Then, in comparison with the other two indices, Xu-index [36] and CH-index [37], three indices of automatic keyword categorization were introduced.

In our study, we defined three sum-of-squares-based indices including the WB-index, CH-index, and Xu-index for determining the number of the cluster. This notion originates from Zhao's [35] research. Before calculating these indices, we need to calculate two basic elements for SSW and SSB. SSW elements are used to measure the compactness of clusters, and SSB elements are used to measure separation. They are defined as

$$\mathrm{SSW(M)} = \max_{t}\left\{\max_{i,j}(1 - IED(T_i, T_j)_{T_i \neq T_j \in C_t})\right\} + \sum_{|C_t=1|} 1 \tag{6}$$

$$\mathrm{SSB(M)} = \sum_{t=1}^{M}\sum_{s>t}^{M} \min\left(1 - IED(T_i, T_j)_{T_i \in C_t, T_j \in C_s}\right) \tag{7}$$

In the SSW formula, T_i and T_j (i, j = {1, 2, . . . , N}) are the ith and jth trajectories in cluster C_t (t = {1, 2, . . . , M}). When only one trajectory is in a cluster, we sum up the number of clusters according to $\sum_{|C_t=1|} 1$. When the number of trajectories is greater than 1, we search the minimum value of edit distance between T_i and T_j. In the SSB formula, C_t and C_s (t, s = {1, 2, . . . , M}) are the tth and sth clusters. T_i and T_j (i, j = {1, 2, . . . , N}) are the ith and jth trajectories in cluster C_t and cluster C_s, respectively. We calculate the cumulative sum for the maximum value of edit distance between T_i and T_j. M is the number of clusters. Therefore, the three indices are defined as

$$\text{WB-index} = M * SSW(M)/SSB(M) \tag{8}$$

$$\text{CH-index} = \frac{SSB(M)/M - 1}{SSW(M)/N - M} \tag{9}$$

$$\text{Xu-index} = \log\sqrt{(SSW(M))/N^2} + logM \tag{10}$$

We can see that the three indices change with the number of clusters using the obtained Formulas (6)–(10). The number of clusters (on the x-axis) is plotted against the three index values

(on the y-axis). By comparing the changes in the three curves, we can determine the optimal number of clusters based on the minimum or maximum values.

3.3. Anomalous Trajectories and Behavior Pattern Detection

3.3.1. Definition of Anomalous Trajectories

We suppose that there are two points: source point (S) and destination point (D). A list of N trajectories $T = \{T_1, T_2, \ldots, T_N\}$ crossing the same SD pairs obtains a set of clustering results $C = \{C_1, C_2, \ldots, C_M\}$. The cluster results include five clusters using the hierarchical clustering method as introduced in Figure 2. The two clusters C_1 and C_2 each contain more than one trajectory. C_1 and C_2 are defined as normal clusters, and these trajectories are defined as normal trajectories, which represent the regular routes of taxi drivers as shown by the gray lines. The other three clusters, C_3, C_4, and C_5, include only one trajectory. C_3, C_4, and C_5 are defined as anomalous clusters, and these trajectories are defined as anomalous trajectories (C_3-t1, C_4-t2, and C_5-t3.), which represent the occasional routes of taxi drivers and are shown in black lines.

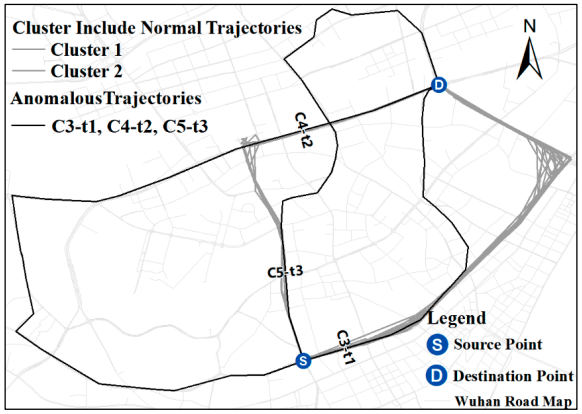


Figure 2. Illustration of all the taxi trajectories between source and destination.

3.3.2. Algorithm Flowchart of Anomalous Trajectory Detection

Our algorithm is presented in a structured pseudocode form of trajectory clustering for anomalous trajectory detection as shown Algorithm 1. It mainly includes three steps and executes two algorithms to perform the tasks, including the edit distance and hierarchical clustering, which have already been introduced in Sections 3.1 and 3.2. For step 1, the process of calculating the edit distance uses dynamic programming techniques and returns the edit distance matrix ED including all trajectories. In step 2, each trajectory forms an initial cluster, and when the distance of two trajectories is minimal, the two trajectories are merged into one cluster until the number of clusters is equal to x. The method of calculating x was introduced in Section 3.2, and its use will be described in detail in Section 4.3. For the final step, we will obtain trajectory data using anomalous attribute labels.

Algorithm 1: Trajectory clustering for anomalous trajectory detection

Input: a trajectory dataset $\boldsymbol{T} = [T_1, T_2, \ldots\ldots, T_N]$; $\boldsymbol{x}$: the optimal number of clusters.
Output: the trajectory clusters $\boldsymbol{C} = [C_1, C_2, \ldots\ldots, C_x]$ and $\boldsymbol{T}$ with anomalous attributes label
Step1: Similarity measurement of trajectories based on improved edit distance algorithm
begin
$l_i (i = 1,2, \ldots N)$: the number of points of T_i; b_{ij}: the normalization value of the edit distance between T_i and T_j; M_{ij}: the edit distance matrix of T_i and T_j; ED: the distance matrix of T.
for $i = 0$ **to** N − 1
  for $j = 0$ **to** i
    pick a pair of trajectories T_i and T_j
    compute the value of the first row and first column of the M_{ij}
      for $i = 2$ to l_i+1
        for $j = 2$ to l_j+1
        min($M_{ij}[i][j]$+insert, $M_{ij}[i][j]$+delete, $M_{ij}[i][j]$+replace) by formula(2)-(4)
      $b_{ij} = M_{ij}[i][j]/(M_{ij}[l_i][0] + M_{ij}[0][l_j])$
      ED[i][j] = b_{ij}
return the edit distance matrix of trajectory dataset ED[i][j]
Step 2: Clustering trajectories into groups by using hierarchical clustering
each trajectory forms an initial cluster C_i=[T_i(i = 1, 2, …, N)], x is the optimal clustering number
repeat
  find the pair of clusters of minimum dissimilarity:
    $d(C_q, C_r) = \arg \min_{T_i \in C_q, T_j \in C_r} ED(T_i, T_j)$
    add $C' = C_q \cup C_r$ to C and delete C_q, C_r from C
    N = N – 1
until N = x
end
return the result of hierarchical clustering $\boldsymbol{C} = [C_1, C_2, \ldots\ldots, C_x]$
Step 3: Label anomalous trajectories based on clustering results
repeat
  find C_i include the number of trajectories only one and label anomalous trajectories
end
return a trajectory dataset $\boldsymbol{T} = [T_1, T_2, \ldots\ldots, T_N]$ with anomalous attributes label

3.3.3. Discovering Anomalous Behavior Patterns Based on Statistical Indicators

The taxi trajectory data only represent the activities of the taxi drivers or passengers [38]. These trajectories reflect the activity processes of taxi drivers or passengers in an urban road network. Usually, normal trajectories included in normal clusters will be chosen by the driver or passengers; however, anomalous trajectories are selected for various reasons, which can be summarized in two subjective or objective perspectives. The first is a subjective perspective including the intentional choice and well-meaning choice of drivers. The intentional choice of drivers is to increase income, and the well-meaning choice of drivers is related to the requirements of passengers. All choices will lead to a significant increase in the lengths of the trajectories. The second is an objective perspective, because unexpected events can happen on the road, causing the trajectory to become anomalous. Unexpected events are accidents, road congestion, road closures, and other special events that result in a significant increase in the time of the trajectories.

There is no prior knowledge of the reasons for the anomalous trajectories. It is feasible to speculate on the reasons based on the statistical properties of anomalous trajectories. Therefore, we chose two appropriate statistical indicators: length and time. Suppose the length and time of an anomalous trajectory A_r are represented by AL_r and AT_r $(r = 1 \cdots n)$, and the average length and time of a normal trajectory are denoted by NL_{value} and NT_{value}. These conditions can be categorized into four kinds of anomalous behavior patterns:

- Anomalous behavior pattern 1(Abp1): $AL_r \leq NL_{value} + L_\rho \; and \; AT_r \leq NT_{value} + T_\rho$,
- Anomalous behavior pattern 2(Abp2): $AL_r \leq NL_{value} + L_\rho \; and \; AT_r > NT_{value} + T_\rho$,
- Anomalous behavior pattern 3(Abp3): $AL_r > NL_{value} + L_\rho \; and \; AT_r \leq NT_{value} + T_\rho$, and
- Anomalous behavior pattern 4(Abp4): $AL_r > NL_{value} + L_\rho \; and \; AT_r > NT_{value} + T_\rho$.

4. Experiment of the Proposed Approach

4.1. Taxi Trajectory Data Pre-Processing

The raw taxi trajectory data include nearly 20,000 taxis and come from a local company in Wuhan City, China in 2014. This dataset includes location and other attribute information—such as speed, direction, state, etc.—recorded at least once every 60 s. Table 1 provides a sample record. Longitudes and latitudes are shown '****' for protecting privacy. The filed 'Acc' represents the state of the engine, which include 'On' and 'OFF' values. The 'On' value means that the engine is working, and the 'OFF' value means that the engine is flameout. The filed 'State' represents whether there are passengers in the taxi, which include 'heave' and 'empty' values. The 'heave' value means that the taxi has passengers, and the 'empty' value means that there is no passenger in the taxi. The taxi trajectory data only represent the activity of a driver or a passenger. However, taxis drivers are more flexible and able to plan their cruising routes, which will form predictable patterns [39]. Raw taxi trajectory data must undergo data filtering, which mainly includes two aspects. In the first aspect, the drift of the GPS device will cause the incorrect recording of the point. These error points need to be removed according to the distances between these points, which are far beyond the distance that a taxi can drive in one minute. In the second aspect, a human or machine generated value errors occurred in the course of recording the text attributes of the trajectories. These errors can be removed through experience, for example, when value of 'State' is 'heave', the value of 'Acc' must be 'On'. Those records that value of 'Acc' is 'OFF' are errors.

Table 1. Sample records from taxi trajectory data.

Vehicle ID	Time	Longitude	Latitude	Direction	Acc	State
1681	00:00:00	114.****	30.****	142	On	empty
7864	00:00:50	114.****	30.****	0	On	heave
...	...	...	...	...	...	...
1681	00:00:50	114.****	30.****	20	On	heave
7864	00:00:60	114.****	30.****	40	On	heave

The data pre-processing involved in the study mainly includes map matching and passenger trajectory extraction of source–destination (SD) pairs. First, the GPS drift problem will lead to inaccurate positioning of the taxi trajectory data. Thus, matching taxi trajectory data to the current road network is necessary. The map matching method is based on the nearest principle where point data for the nearest road factor are matched within a certain search radius. Second, given a pair of SD points, we need to select all driving trajectories between the SD pair. The driving trajectories are those trajectories crossing the same SD pair, and their values of 'State' is 'heavy'.

In our experiment, four source–destination pairs (SD1, SD2, SD3, and SD4) are selected, and each (SD) pair represents an actual location (Table 2). To obtain more trajectories, these locations include railway stations, business circles, parks, etc., where people have frequent activities.

Table 2. Geographic names of four source–destination group.

Group Name	First Group (SD1)	Second Group (SD2)	Third Group (SD3)	Fourth Group (SD4)
Source (S)	JieDaoKou	ZhongJian KangCheng	ZhongJian KangCheng	Wuchang Railway Station
Destination (D)	Wuhan Railway Station	MaAnShan Forest Park	JieDaoKou	WuHan Railway Station

All driving trajectories, including the number of trajectories between each SD pair, are overlaid with a Wuhan city road map, as shown in Figure 3.

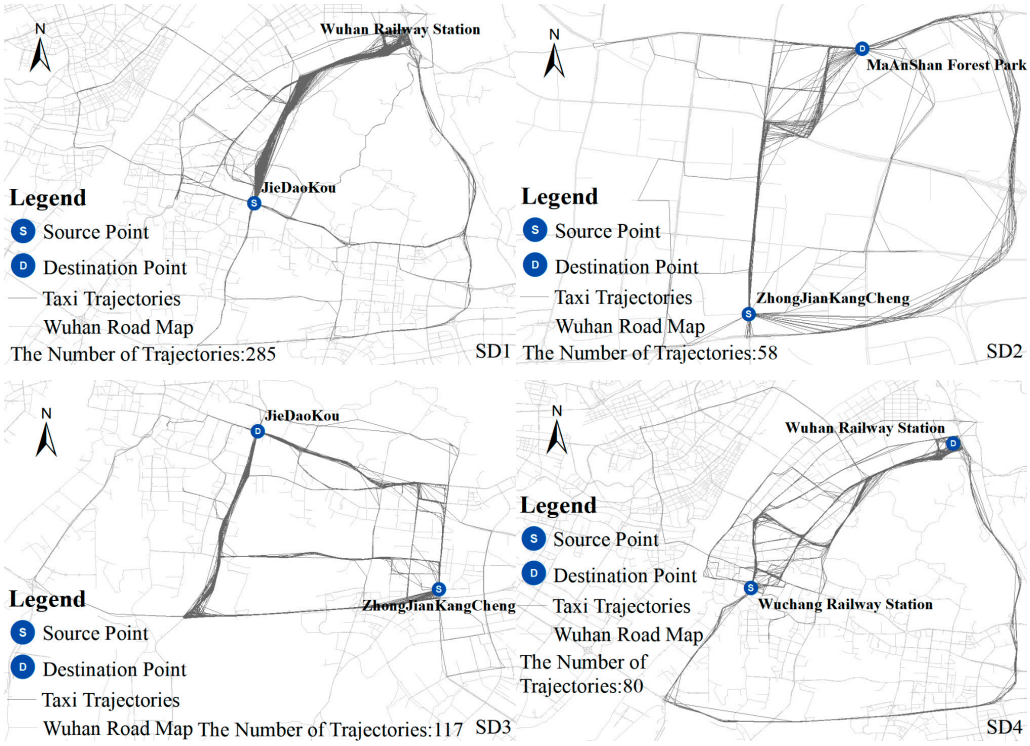


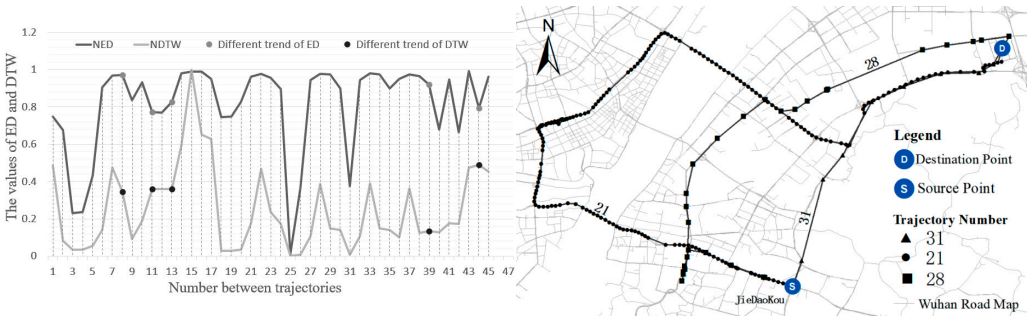
Figure 3. Geographic distribution of four SD pairs of experimental data.

4.2. Distance Measurement of Taxi Trajectory Data

Based on the proposed improved edit distance algorithm (Section 3.1), we randomly selected 10 real taxi passenger trajectories from SD1 data. We calculated the edit distance between the trajectories and obtained the 45 values each two trajectories. To prove the validity of the edit distance method, the DTW (dynamic time warping) [40] method was chosen for comparison. The key feature of DTW is that it allows siftings and elongations while it compares two time-series, which is a common method in time series data [41], and can be used for trajectory data. A comparison of the IED and DTW values is shown in Figure 4a for 10 trajectory datasets. To show convenience, DTW normalization is based on its maximum and minimum values.

We know that the changing trends of the values of IED and DTW are roughly the same as seen in Figure 4a. However, the different trends of values for IED and DTW are marked with the numbers 8, 11, 13, 39, and 44. To identify the reasons, we chose the corresponding trajectories for the marked number and its previous number. All results of IED and DTW from these trajectories are shown in Table 3. NIED and NDTW are the values of IED and DTW standardization. From the overall values of this view, it is found that the most involved trajectories are at 21. Therefore, we take the example and choose two numbers, 10 and 11, corresponding to the trajectory IDs of 21 and 28 as well as 21 and

31 from Table 3, respectively. We know that the IED values of trajectories 21 and 28 are greater than those of trajectories 21 and 31, which means that the trajectories 21 and 31 are more similar. In contrast, the DTW values of trajectories 21 and 28 are smaller than those of trajectories 21 and 31, which means that trajectories 21 and 28 are more similar. The spatial distribution of trajectories 21, 28, and 31 in Wuhan are shown in Figure 4b.



(**a**) (**b**)

Figure 4. (**a**) Comparing between ED and DTW of trajectories; (**b**) Geographic distribution of compared three trajectories 21, 28, and 31.

Table 3. The values of ED and DTW between trajectories.

Number	Trajectory ID Pairs	IED	NIED	DTW	NDTW
07	1, 180	63.6384	0.9681	1538.5648	0.4743
08	1, 188	38.1715	0.9723	1115.1115	0.3437
10	21, 28	61.0285	0.9325	612.7917	0.1888
11	21, 31	43.2647	0.7715	1166.6670	0.3596
12	21, 68	43.1681	0.7700	1168.2111	0.3601
13	21, 82	49.4162	0.82741	1165.6056	0.3593
38	21, 126	40.5420	0.9645	401.3309	0.1235
39	21, 120	40.0110	0.9209	438.9380	0.1351
43	82, 188	75.7054	0.9913	1540.9585	0.4750
44	82, 209	61.7448	0.7938	1587.8943	0.4895

Trajectories 21 and 31 are more similar than trajectories 21 and 28 from the spatial distribution of the trajectories as seen on the map. Because the trajectory is recorded in time intervals of less than 60 s, the sampling rate is not fixed, which will lead to a large number of sampling points in some trajectories, for example, the number of points included in trajectory 21 is 217; however, the time of trajectory 21 is 43.33 min, which is far longer than a minute to sample each point. When we calculate trajectories 21 and 31 using the DTW method, some of the sampling points in trajectory 21 will be used many times to achieve local stretching of the time dimension. This will result in an inaccurate calculation of the DTW value, which is due to the inconsistency of the sampling rate for the trajectory data. Our IED method can overcome such a problem according to standardization to realize spatial dimension scaling. Therefore, the edit distance method can address the problem of inaccurate calculation of distance values caused by the large disparities in sampling rates for the trajectories.

4.3. Comparison of Indices for Automatic Trajectory Clustering

Four SD pairs of trajectory data are involved in the experiment. According to Section 3.2, we calculate three sum-of-squares-based indices—the WB-index, CH-index, and Xu-index—for automatic trajectory clustering. The distance between trajectories is based on an improved edit distance algorithm as described in Section 3.1. The hierarchical clustering method is used to cluster trajectories into groups in the progress of calculating indices. For our experimental data, the number of clusters over 100 is meaningless. Therefore, the maximum number of clusters is 100.

The number of clusters determined by the three sum-of-squares-based indices on the four SD pairs of trajectory data are shown in Figure 5. The change in the Xu-index index is monotonically increasing; thus, this index cannot be used to determine the number of clusters. For SD2, SD3, and SD4, there is a clear maximum value for the CH-index at 8, 20, and 24. The minimum values of the WB-index are also at 8, 20, and 24. For SD1, the maximum value of the CH-index is at 21; however, the minimum values of the WB-index are also at 20. The numbers of clusters corresponding to these values are inconsistent. This requires human judgement to determine which is best by comparing the clustering result. Of course, when the number of clusters is very close, for SD1, we are not good at making judgments from the result. When the data size is large, the factor $(M-1)/(N-m)$ plays a more important role than SSW/SSB in the whole index [35]. Therefore, when the values of the WB-index and CH-index are different, we chose the maximum of the CH-index as the number of clusters.

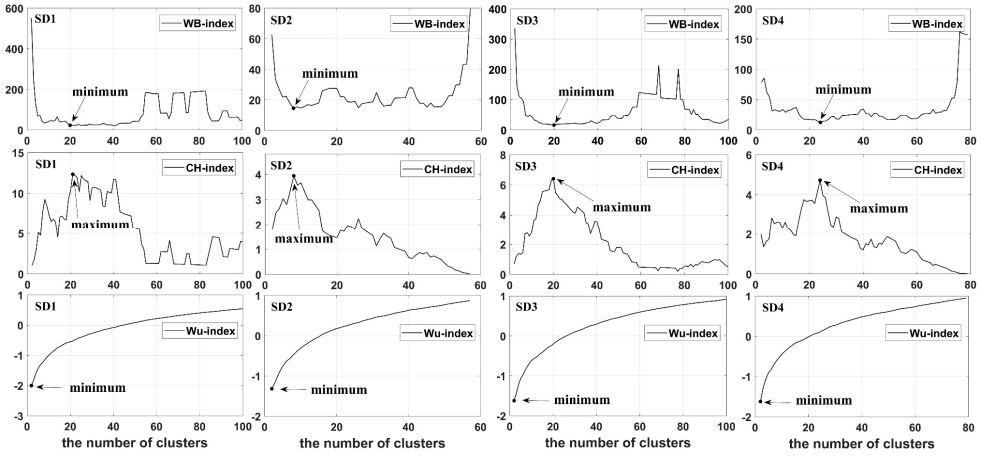


Figure 5. Relation graph for the number of clusters and three sum-of-squares based indices.

4.4. Anomalous Trajectory Detection by Trajectory Clustering

We can choose the appropriate linkage methods through experiments by using test data. The test data also are from the real taxi trajectory data and were already introduced in Section 3.3 (shown in Figure 2), with an obvious five clusters and two anomalous trajectories. The number of clusters is set to 5. The clustering results based on different linkage methods—which include single linkage (S), complete linkage (C1), average linkage (A), Ward's method (W), and the centroid method (C2)—are shown in Table 4. The IDs of the anomalous trajectories are t1, t2, and t3. Only the S method detected the three anomalous trajectories correctly. The C1 and W methods identified two anomalous trajectories, and the A and C2 methods only found one anomalous trajectory.

Table 4. The clustering result based on different linkage method.

ID	t1	2	3	4	t2	6	7	8	9	10	11	12	13	14	15	16
S	0	4	4	4	1	4	4	4	4	4	3	4	4	4	3	4
C1	3	2	2	2	0	2	3	2	3	3	4	3	2	2	1	3
A	4	2	2	2	0	2	4	2	4	4	3	4	2	2	3	4
W	3	2	2	2	0	2	1	2	1	1	4	1	2	2	4	1
C2	0	2	2	2	1	2	3	2	3	3	4	3	2	2	4	3
ID	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**t3**	**27**	**28**	**29**	**30**	**31**	
S	4	4	4	4	3	4	4	3	4	2	4	4	3	4	4	
C1	3	3	2	3	4	2	3	4	2	4	2	2	1	2	3	
A	4	4	2	4	3	2	4	3	2	1	2	2	3	2	4	
W	1	3	2	1	4	2	3	4	2	4	2	2	4	2	1	
C2	3	3	2	3	4	2	3	4	2	4	2	2	4	2	3	

Complete linkage is suitable for determining a relatively compact cluster. Average linkage considers the structure of the class, which is suitable for two classes of small difference. Furthermore, unlike the distance between points, in our analysis, the smaller the distance, the more similar the trajectories. The centroid of taxi trajectories is different from call data based on Yuan's [28] method. Taxi trajectory data are restricted by road networks, and the centroid of the trajectory by calculating the average position is not on the roads, which does not fully represent the trajectory and deviates from reality significance. Thus, the centroid method cannot be used directly and needs to be improved. Ward's method [32] also belongs to the centroid method and is suitable for datasets in which the number of data in each cluster is approximately equal and has no anomalous value. Meanwhile, Ward's method needs to calculate an average object in a cluster, which leads to the same problem as the centroid method. Therefore, we chose the single linkage method as the linkage between clusters by experiment and analysis.

According to the Section 4.3, the optimal cluster numbers for SD1, SD2, SD3, and SD4 are 21, 8, 20, and 24, respectively. Clustering results can be overlapped with the Wuhan road map overlay, as shown in Figure 6. To show the clustering results clearly, we display the clustering results of each SD1–SD4 pair in four sub-figures. The first three sub-figures display the normal clustering and are represented by a colored solid line, and the number of clusters displayed in each sub-figure is different. In addition, the average length and time of the trajectories in the normal clusters are shown. The last sub-figure shows the anomalous trajectories, which are denoted by a black dotted line, and the average length and time of the anomalous trajectories are also shown. The Wuhan road map is indicated by a gray line. The units of length and time are kilometers and minutes, respectively.

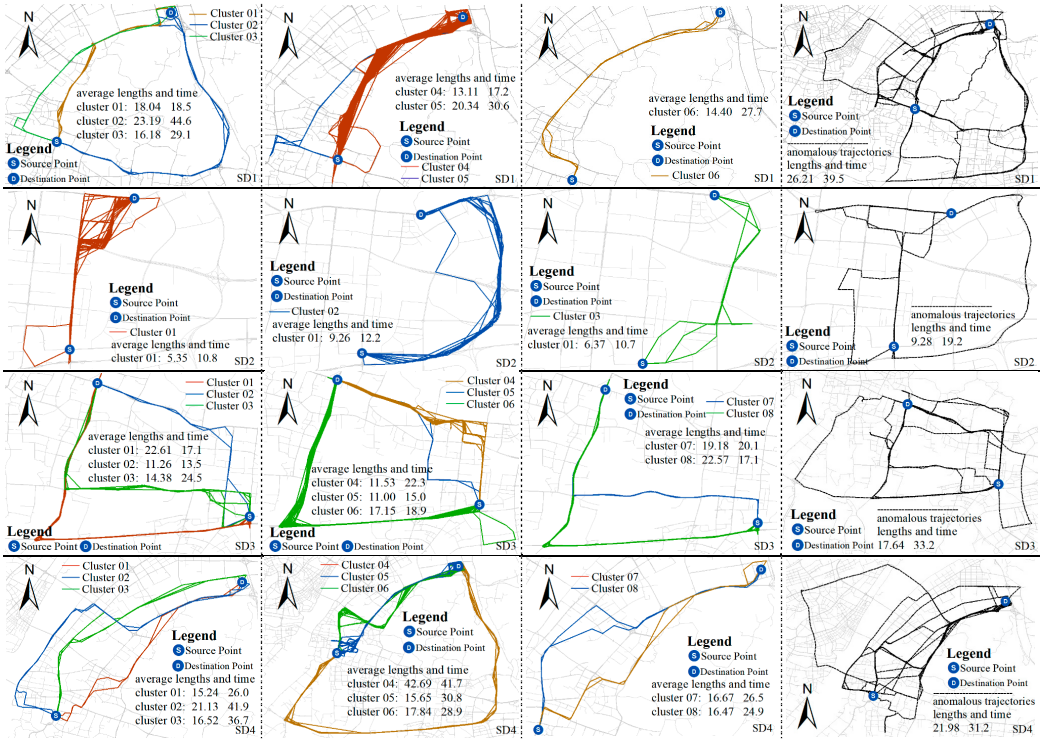


Figure 6. Geographic distribution of clustering results in SD1–SD4 pairs of experimental data.

4.5. Anomalous Trajectory Behavior Pattern Analysis

According to Section 3.3, we also need to set threshold L_ρ and time threshold T_ρ to determine the behavior patterns. According to the distribution of the SD1–SD4 data (shown in Figure 7), the distribution of length and time belonged to a long tail distribution. In Wuhan city, the price for a taxi ride is 10 yuan for the first three kilometers and then an additional 1.8 yuan per kilometer.

In most cases, the driver chooses to increase the length of the route using a detour to augment income. According to Figure 7, the minimum length of SD1–SD4 data is 4.83 kilometers; thus, the length threshold L_ρ set to five kilometers is reasonable. Meanwhile, the minimum length of SD1–SD4 data is 6.98 min, and time increased to five minutes becomes a barrier to travel for passengers. Therefore, the time threshold T_ρ is set to five minutes.

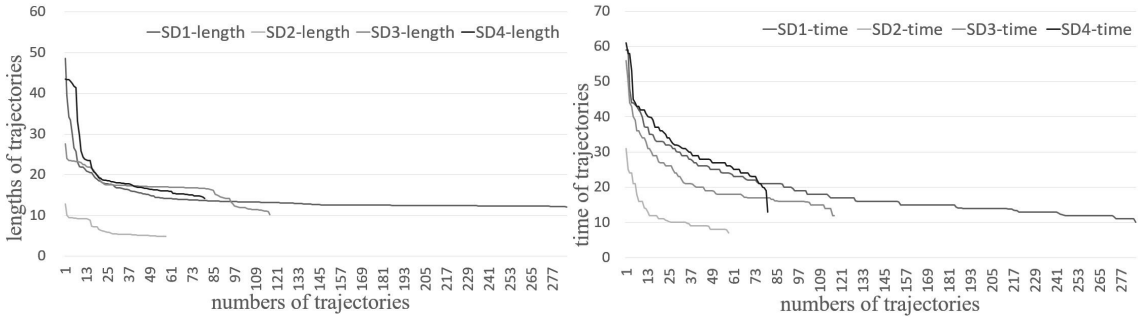


Figure 7. Lengths and time distribution of the SD1–SD4 data.

The statistical results of length and time for the trajectories from SD1–SD4 data are shown in Figure 8. It is a 2D plot, where x is the length of the anomalous trajectories and y is the time of the anomalous trajectories. Vertical and horizontal lines, respectively, are the average length and time of normal trajectories.

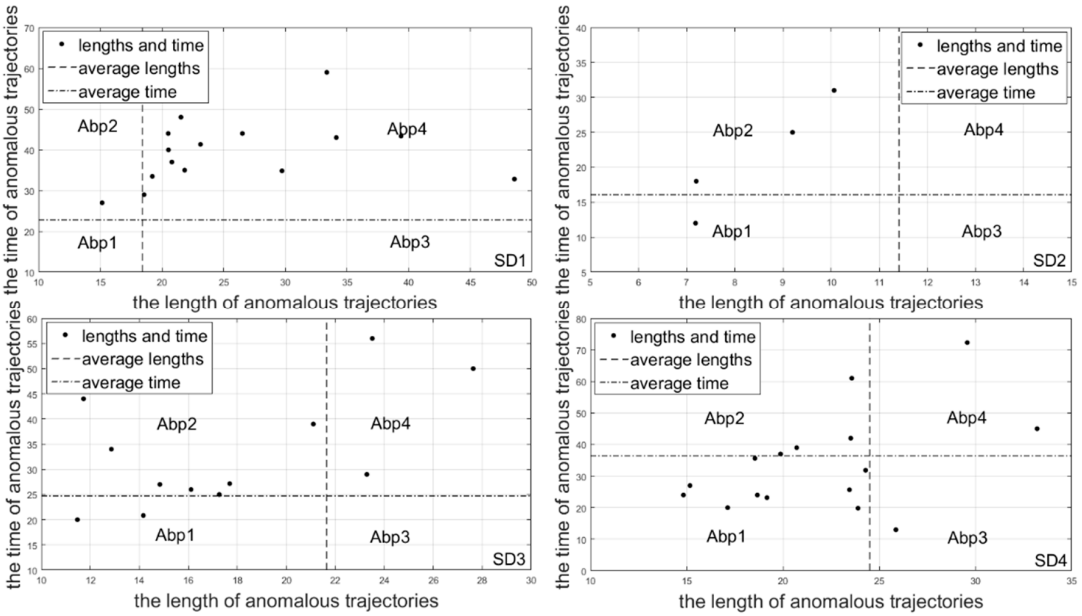


Figure 8. Classification graph on anomalous trajectories lengths and time.

By statistical classification from Figure 8, we can see that the SD1 data contain two anomalous behavior patterns (Abp2 and Abp4), the SD2 data contain two anomalous behavior patterns (Abp1 and Abp2), the SD3 data contain three anomalous behavior patterns (Abp1, Abp2, and Abp4), and the SD4 data contain four anomalous behavior patterns (Abp1, Abp2, Abp3, and Abp4). The geographic distributions of the four behavior patterns are represented in Figure 9. The four behavior patterns are expressed using different colors. The changes in color depth indicate the times of the trajectories; light-colored trajectories are relatively short, and dark-colored trajectories are relatively long. The lengths of the trajectories are tagged on the map.

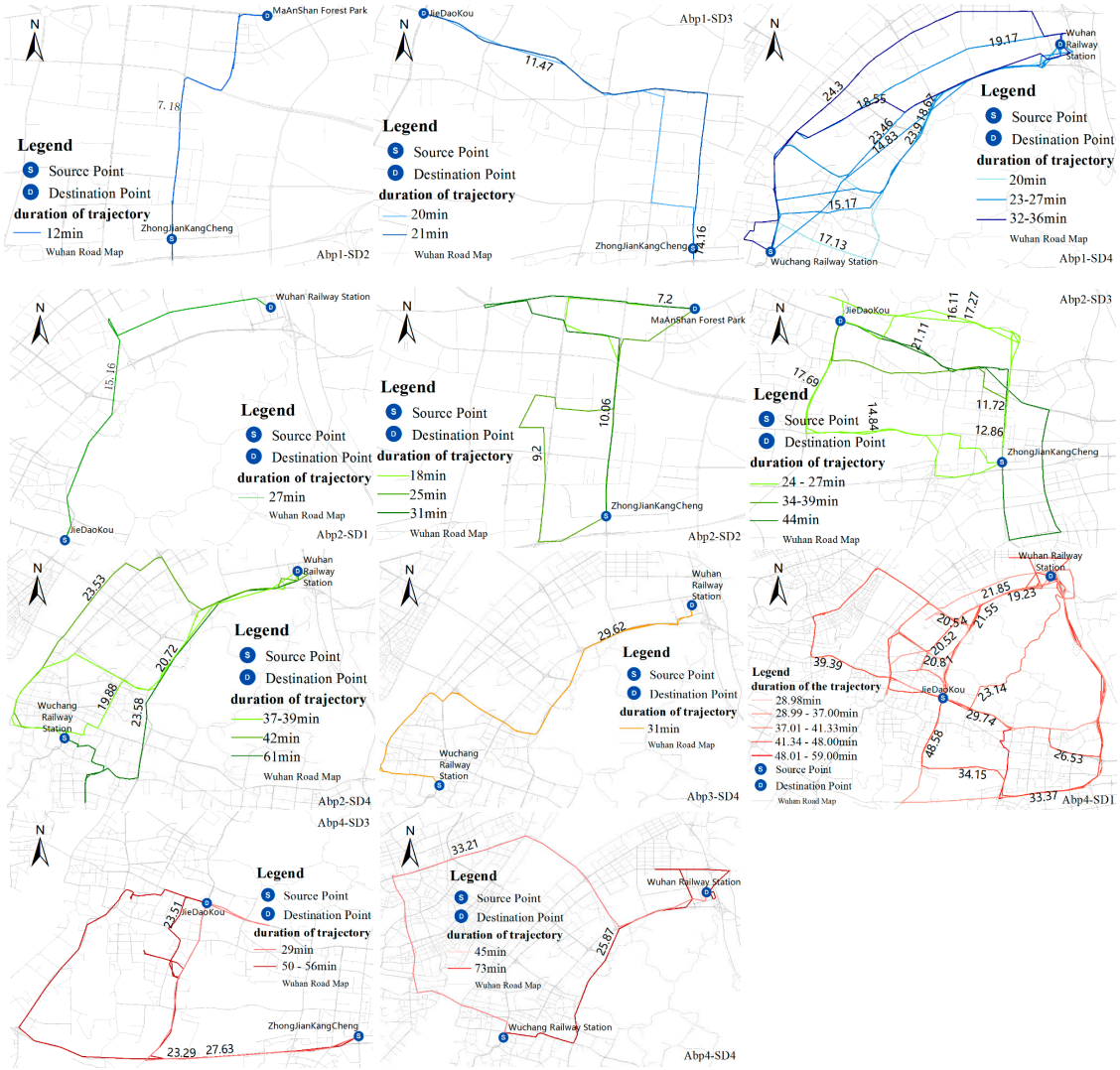


Figure 9. Geographic distribution of behavior patterns in the four pairs of experimental data.

In Abp1, there are many anomalous trajectories that occur in the early hours of the morning, including all trajectories in Abp1-SD3 and the light-colored trajectories in Abp1-SD4. These time periods represent travel on uncongested roads and conform to the characteristics of the trajectory. Moreover, in Abp1-SD2, the occurrence of anomalous trajectories is between 11:45–11:57, and the dark-colored trajectories in Abp1-SD4 also occur between 12:37–13:52. These time periods represent rush hour in the city, when terrible traffic can be expected. However, the lengths and times of these trajectories are relatively small. This can be explained by a taxi driver choosing an uncongested road to avoid congested roads during rush hour based on experience. Thus, Abp1 can be used as a recommended route for residents during rush hour from S to D.

In Abp2, many anomalous trajectories are similar to the normal cluster, such as Abp2-SD1 being similar to Cluster01 in SD1, the light-colored trajectories in Abp2-SD3 similar to Cluster03 and Cluster04 in SD3, and the dark-colored trajectories in Abp2-SD4 similar to Cluster02 in SD4. However, the time of these trajectories is greater than the average time of these normal clusters. Meanwhile, many anomalous trajectories occur at rush hour. This indicates that some unusual events are most likely to occur at this time, such as congestion, traffic accidents, etc. Thus, we can determine whether unexpected events occur on the road based on the time and place of anomalous trajectories belonging to Abp2 combined with the Wuhan road map.

In Abp3, there was only one trajectory from Abp3-SD4. The starting point and the destination are all the train stations in the SD4, the driver chooses the long route but takes less time to reach the destination, and the occurrence of the anomalous trajectories is between 05:09–05:41. This can be explained by the driver selecting the route with no congestion to increase speed and reach the destination in a timely manner, as passengers may need to hurry to another station in the early hours of the morning. Therefore, Abp3 can be used to identify a driver who chose a route with a detour to help passengers arrive at their destination more quickly.

In Abp4, most of the trajectories are more than 30 min in time, and most of the trajectories are greater than 20 kilometers in length. These can be identified as significant detour trajectories. For example, in Abp4-SD3, the three trajectories all choose left first through the three ring roads and then the destination. One of them even travels to the left to the edge of the Yangtze River and back to the north. In Abp4- SD4, one anomalous trajectory chooses to pass over the Yangtze River twice from the Wuchang Railway Station to the Wuhan Railway Station. The width of the Yangtze River through Wuhan city is approximately 1.4 km. This can be explained by evident detour behavior. However, we cannot be certain that these trajectories are the driver's fraudulent behavior without a priori knowledge. It is also possible to go to one place first and then finally to the destination because of the passengers' requirements.

5. Discussion

The similarities of taxi GPS trajectories can be measured effectively using an improved edit distance method according to the characteristics of the trajectory data. We compare two common methods include DTW and IED methods in Section 4.2. The DTW method calculates distance by stretching or scaling the time dimension, which ensures that the time sequence of the trajectory record points is the same, and does not need to be compared in one-to-one time. However, DTW is robust in response to an increase in the sampling rate but highly sensitive to a decrease [42]. Therefore, the DTW method can lead to some inaccurate calculations when the difference in the sampling rate of the trajectories is relatively large (See Section 4.2). A trajectory is a group of limited points in a time sequence that can be divided into n sub-intervals consisting of corresponding points. The IED method can search for the number of similar sub-intervals between two trajectories. This method does not require the sequences of the two trajectories to have a point-by-point correspondence relationship. Instead, the method can better reflect the structural difference of the trajectory sequences and determine the non-overall similar trajectories.

Taxi anomalous trajectories are some driving trajectories chosen by a small number of drivers that are different from the regular choices of other drivers. The definitions of anomalous trajectories and normal trajectories are introduced in Section 3.3.1, and these are suitable for detection using the hierarchical-clustering method. Anomalous trajectories are equivalent to the isolated leaf nodes in the hierarchical tree, the stem length represents the distance between the trajectories, and the isolated leaf nodes and adjacent layers have a longer stem length. The only parameter of hierarchical clustering is the number of clusters, which is determined based on comparing three sum-of-squares-based indices: the WB-index, CH-index, and Xu-index. For our trajectory data, the CH-index is more effective compared to the WB-index and XU-index. When the number of trajectories is large, these indicators cannot provide the true number of clusters by comparing the clustering result with those of human judgement. However, in our experiment, these indices are helpful for using the clustering method.

The time complexity for the edit distance method depends on the number of trajectories and the size of the database. Suppose that these have N list of trajectories $T = \{T_1, T_2, \ldots, T_N\}$. For the trajectory T_i, the numbers T_i and T_j are m and n, therefore, the time complexity of calculating the distance between T_i and T_j is O(m*n), and the time complexity of calculating N trajectories is $O(N^2)$. The trajectories are independent of each other. When the amount of data is large, we consider dividing the trajectories into multiple groups and calculating the distance between them in parallel. Usually, the time complexity of hierarchical clustering is $O(N^2)$, and the space complexity of hierarchical

clustering is $O(N^3)$. To reduce the computational complexity of hierarchical clustering, we chose the nearest-neighbor chain algorithm [43]. In this algorithm, the nearest neighbor chain is used to decide the merging of these clusters. Each cluster in the nearest neighbor chain is the nearest neighbor of its nearest cluster, that is, the cluster with the smallest distance. The algorithm can reduce the time complexity of hierarchical clustering to $O(N^2)$.

The aim of this paper is to detect anomalous trajectories based on trajectory clustering method. The analysis of anomalous trajectory behavior patterns can help more effectively explain the cause and significance of these anomalous trajectories. The length and time of their distribution belonged to a long-tail distribution as seen in Figure 7. We sort the length and time respectively and obtain the trajectory IDs corresponding to the 20 percent larger values of the time and the length from SD1–SD4. They do not completely contain the trajectory IDs of the anomalous trajectories. Especially anomalous trajectory IDs belong to Abp4. Certainly, simple length and time attributes cannot distinguish the similarity between these trajectories in geographic space. To distinguish them, more statistical indicators need to be added. However, this requires more statistical analysis and is not the focus of this study.

Unavoidable uncertainty issues also exist when detecting anomalous trajectories utilizing taxi trajectory data. In our analysis, uncertainty issues need to be discussed mainly in three aspects. The first aspect is data quality. Taxi position may not be accurate because of the drift in GPS data. Uncertainty in passenger trajectory could also be caused by lower frequencies and inaccurate sampling points. The second aspect refers to the uncertainty of the model and algorithm. Although we can prove the effectiveness of the method in anomalous trajectory detection, selecting different methods will lead to uncertainty in the results. Examples include different clustering methods or clustering number determination methods. The final aspect is mobile randomness of the driver. Although anomalous behaviors can be analyzed, random and distinct characteristics of driver behaviors are inevitable, which can lead to uncertainty about some anomalous behaviors without prior knowledge.

6. Conclusions and Future Research

In this paper, we proposed a trajectory clustering method to detect anomalous trajectories. We improved the operation cost of the edit distance algorithm, and we used this method to obtain similarity measurements of taxi GPS trajectories. During the hierarchical clustering process, we determined the number of clusters based on a comparison of three sum-of-squares-based indices. According to the algorithm flowchart of the anomalous trajectory detection, we detect anomalous trajectories and then analyze anomalous behavior patterns according to the length and time of the anomalous trajectories. The main contribution of this paper is to consider the trajectory's integrality in discovering anomalous taxi trajectories using a trajectory clustering method that can effectively perform automatic mode detection using taxi GPS trajectories. Moreover, this paper detects four anomalous behavior patterns and summarizes the causes of the behavior of the anomalous trajectory. It is meaningful to discover the unusual behavior of taxi drivers and anomalous traffic conditions.

Our future research includes the following aspects: (1) Current clustering objects are driving trajectories that confined same SD pairs. In fact, the source point and destination point of each driving trajectory is different. In the future, taking the trajectories of different SD pairs as the research object, we study a suitable method to detect the anomalous trajectories. (2) This paper employed historical taxi trajectory data. However, some anomalous behavior patterns need to be detected and processed in real time. In future research, we will detect anomalous behavior of taxis based on real-time taxi GPS trajectory data and carry out detection and analysis of anomalies online, which will further improve the effectiveness of the anomalous trajectory analysis. (3) Four anomalous behavior patterns were discovered based on the clustering results of the statistical analysis. In the future, we need to thoroughly analyze the causes of anomalous behavior from the perspectives of human behavior and dynamic city development, making the categorization of anomalous behavior more reasonable.

Acknowledgments: This research was funded by the National Key Research and Development Program (no. 2017YFB0503604) and the National Natural Science Foundation of China (no. 41471326).

Author Contributions: Yulong Wang and Kun Qin conceived and designed the anomalous trajectory detecting algorithms in this paper. Yulong Wang performed the experiments and wrote the paper with Kun Qin together. Yixiang Chen contributed result analysis and discussed the idea. Pengxiang Zhao revised and discussed the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, S.; Wang, Z. Correction: Inferring Passenger Denial Behavior of Taxi Drivers from Large-Scale Taxi Traces. *PLoS ONE* **2017**, *12*, e0171876. [CrossRef] [PubMed]
2. Liu, X.; Ban, Y. Uncovering Spatio-Temporal Cluster Patterns Using Massive Floating Car Data. *ISPRS Int. J. Geo-Inf.* **2013**, *2*, 371–384. [CrossRef]
3. Yin, P.; Ye, M.; Lee, W.C.; Li, Z. Mining GPS Data for Trajectory Recommendation. In *Advances in Knowledge Discovery and Data Mining*; Springer International Publishing: Cham, Switzerland, 2014; pp. 50–61.
4. Matsubara, Y.; Li, L.; Papalexakis, E.; Lo, D.; Sakurai, Y.; Faloutsos, C. F-Trail: Finding Patterns in Taxi Trajectories. In *Advances in Knowledge Discovery and Data Mining*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 86–98.
5. Liu, S.; Ni, L.M.; Krishnan, R. Fraud Detection from Taxis' Driving Behaviors. *IEEE Trans. Veh. Technol.* **2014**, *63*, 464–472. [CrossRef]
6. Zheng, Y.; Liu, Y.; Yuan, J.; Xie, X. Urban computing with taxicabs. In *International Conference on Ubiquitous Computing*; ACM: New York, NY, USA, 2011; pp. 89–98.
7. Li, Z.; Filev, D.P.; Kolmanovsky, I.; Atkins, E.; Lu, J. A New Clustering Algorithm for Processing GPS-Based Road Anomaly Reports with a Mahalanobis Distance. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1980–1988. [CrossRef]
8. Chen, Q.; Qiu, Q.; Li, H.; Wu, Q. A neuromorphic architecture for anomaly detection in autonomous large-area traffic monitoring. In Proceedings of the IEEE International Conference on Computer-Aided Design, San Jose, CA, USA, 18–21 November 2013; pp. 202–205.
9. Wang, Z.; Lu, M.; Yuan, X.; Zhang, J.; Van De Wetering, H. Visual Traffic Jam Analysis Based on Trajectory Data. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 2159–2168. [CrossRef] [PubMed]
10. Yuan, J.; Zheng, Y.; Zhang, C.; Xie, W.; Xie, X.; Sun, G.; Huang, Y. T-drive: Driving directions based on taxi trajectories. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; ACM: New York, NY, USA, 2010; pp. 99–108.
11. Ge, Y.; Xiong, H.; Liu, C.; Zhou, Z.H. A Taxi Driving Fraud Detection System. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining, Vancouver, BC, Canada, 11–14 December 2011; pp. 181–190.
12. Zhang, D.; Li, N.; Zhou, Z.H.; Chen, C.; Sun, L.; Li, S. iBAT: Detecting anomalous taxi trajectories from GPS traces. In Proceedings of the 13th international conference on Ubiquitous computing, Beijing, China, 17–21 September 2011; ACM: New York, NY, USA, 2011; pp. 99–108.
13. Kim, J.; Mahmassani, H.S. Spatial and Temporal Characterization of Travel Patterns in a Traffic Network Using Vehicle Trajectories. *Transp. Res. Part C Emerg. Technol.* **2015**, *9*, 164–184.
14. Huang, H. Anomalous behavior detection in single-trajectory data. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 2075–2094. [CrossRef]
15. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; ACM: New York, NY, USA, 2010; pp. 851–860.
16. Kumar, D.; Bezdek, J.C.; Rajasegarar, S.; Leckie, C.; Palaniswami, M. A visual-numeric approach to clustering and anomaly detection for trajectory data. *Vis. Comput. Int. J. Comput. Graph.* **2017**, *33*, 1–17. [CrossRef]
17. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 75–79. [CrossRef]
18. Hwang, J.R.; Kang, H.Y.; Li, K.J. Spatio-Temporal Similarity Analysis between Trajectories on Road Networks. In *Perspectives in Conceptual Modeling*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 280–289.

19. Lee, J.G.; Han, J.; Li, X. Trajectory Outlier Detection: A Partition-and-Detect Framework. In Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, Cancun, Mexico, 7–12 April 2008; pp. 140–149.
20. Guan, B.; Liu, L.; Chen, J. Using Relative Distance and Hausdorff Distance to Mine Trajectory Clusters. *Telkomnika Indones. J. Electr. Eng.* **2013**, *11*, 115–122. [CrossRef]
21. Won, J.I.; Kim, S.W.; Baek, J.H.; Lee, J. Trajectory clustering in road network environment. In Proceedings of the CIDM IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, USA, 30 March–2 April 2009; pp. 299–305.
22. Sha, W.; Xiao, Y.; Wang, H.; Li, Y.; Wang, X. Searching for spatio-temporal similar trajectories on road networks using Network Voronoi Diagram. *Commun. Comput. Inf. Sci.* **2015**, *482*, 361–371.
23. Lee, J.G.; Han, J.; Whang, K.Y. Trajectory clustering: A partition-and-group framework. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, Beijing, China, 11–14 June 2007; ACM: New York, NY, USA, 2007; pp. 593–604.
24. Abraham, S.; Lal, P.S. Spatio-temporal similarity of network-constrained moving object trajectories using sequence alignment of travel locations. *Transp. Res. Part C Emerg. Technol.* **2012**, *23*, 109–123. [CrossRef]
25. Chen, L.; Ng, R. On The Marriage of Lp-norms and Edit Distance. In Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, ON, Canada, 31 August–3 September 2004; pp. 792–803.
26. Chen, L.; Özsu, M.T.; Oria, V. Robust and fast similarity search for moving object trajectories. In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, MD, USA, 14–16 June 2005; pp. 491–502.
27. Dodge, S.; Laube, P.; Weibel, R. Movement similarity assessment using symbolic representation of trajectories. *Int. J. Geogr. Inf. Sci.* **2012**, *26*, 1563–1588. [CrossRef]
28. Yuan, Y.; Raubal, M. Measuring similarity of mobile phone user trajectories–a Spatio-temporal Edit Distance method. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 496–520. [CrossRef]
29. Bermingham, L.; Lee, I. A general methodology for n-dimensional trajectory clustering. *Expert Syst. Appl.* **2015**, *42*, 7573–7581. [CrossRef]
30. Fu, Z.; Hu, W.; Tan, T. Similarity based vehicle trajectory clustering and anomaly detection. In Proceedings of the 2005 IEEE International Conference on Image Processing, Genova, Italy, 14 September 2005.
31. Roh, G.P.; Hwang, S.W. NNCluster: An efficient clustering algorithm for road network trajectories. In Proceedings of the 15th International Conference on Database Systems for Advanced Applications, Tsukuba, Japan, 1–4 April 2010; Springer: Berlin/Heidelberg, Germany, 2010; Volume Part II, pp. 47–61.
32. Amorim, R.C. *Feature Relevance in Ward's Hierarchical Clustering Using the Lp Norm*; Springer: New York, NY, USA, 2015.
33. Ketchen, D.J.; Shook, C.L. The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strateg. Manag. J.* **1996**, *17*, 441–458. [CrossRef]
34. Salvador, S.; Chan, P. Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. In Proceedings of the IEEE International Conference on Tools with Artificial Intelligence, Boca Raton, FL, USA, 15–17 November 2004; pp. 576–584.
35. Zhao, Q.; Fränti, P. WB-index: A sum-of-squares based index for cluster validity. *Data Knowl. Eng.* **2014**, *92*, 77–89. [CrossRef]
36. Xu, L. Bayesian Ying–Yang machine, clustering and number of clusters. *Pattern Recognit. Lett.* **1997**, *18*, 1167–1178. [CrossRef]
37. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **1974**, *3*, 1–27.
38. Zhao, P.; Qin, K.; Ye, X.; Wang, Y.; Chen, Y. A trajectory clustering approach based on decision graph and data field for detecting hotspots. *Int. J. Geogr. Inf. Syst.* **2016**, *31*, 1101–1127. [CrossRef]
39. Wu, L.; Hu, S.; Yin, L.; Wang, Y.; Chen, Z.; Guo, M.; Xie, Z. Optimizing Cruising Routes for Taxi Drivers Using a Spatio-Temporal Trajectory Model. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 373. [CrossRef]
40. Buza, K.A. *Fusion Methods for Time-Series Classification*; Peter Lang: Bern, Switzerland, 2011.
41. Marussy, K.; Buza, K. SUCCESS: A New Approach for Semi-supervised Classification of Time-Series. In *Artificial Intelligence and Soft Computing*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 437–447.

42. Mariescu-Istodor, R.; Fränti, P. Grid-Based Method for GPS Route Analysis for Retrieval. *ACM Trans. Spat. Algorithms Syst.* **2017**, *3*, 1–28. [CrossRef]
43. Murtagh, F. Clustering in Massive Data Sets. In *Handbook of Massive Data Sets*; Abello, J., Pardalos, P.M., Resende, M.G.C., Eds.; Springer: Boston, MA, USA; pp. 501–543.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).