electronics

MDPI

Article

The Fast Detection of Abnormal ETC Data Based on an Improved DTW Algorithm

Feng Guo [1], Fumin Zou [1,2,*], Sijie Luo [2,*], Lyuchao Liao [2], Jinshan Wu [2], Xiang Yu [2] and Cheng Zhang [3]

[1] College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China; n180310004@fzu.edu.cn
[2] Fujian Key Laboratory for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350118, China; achao@fjut.edu.cn (L.L.); 2191905037@smail.fjut.edu.cn (J.W.); 2201901004@smail.fjut.edu.cn (X.Y.)
[3] College of Information Technology and Management, Hunan University of Finance and Economics, Changsha 410205, China; 201805230240@mails.hufe.edu.cn
* Correspondence: fmzou@fjut.edu.cn (F.Z.); sjluo@fjut.edu.cn (S.L.); Tel.: +86-136-3842-8527 (S.L.)

Abstract: As one of the largest Internet of Things systems in the world, China's expressway electronic toll collection (ETC) generates nearly one billion pieces of transaction data every day, recording the traffic trajectories of almost all vehicles on the expressway, which has great potential application value. However, there are inevitable missed transactions and false transactions in the expressway ETC system, which leads to certain false and missing rates in ETC data. In this work, a dynamic search step SegrDTW algorithm based on an improved DTW algorithm is proposed according to the characteristics of expressway ETC data with origin–destination (OD) data constraints and coupling between the gantry path and the vehicle trajectory. Through constructing the spatial window of segment retrieval, the spatial complexity of the DTW algorithm is effectively reduced, and the efficiency of the abnormal ETC data detection is greatly improved. In real traffic data experiments, the SegrDTW algorithm only needs 3.36 s to measure the abnormal events of a single set of OD path data for 10 days. Compared with the mainstream algorithms, the SegrDTW performs best. Therefore, the proposal provides a feasible method for the abnormal event detection of expressway ETC data in a province and even the whole country.

Keywords: data governance; electronic toll collection data; trajectory similarity; dynamic time warping; anomaly detection; expressway

check for updates

Citation: Guo, F.; Zou, F.; Luo, S.; Liao, L.; Wu, J.; Yu, X.; Zhang, C. The Fast Detection of Abnormal ETC Data Based on an Improved DTW Algorithm. *Electronics* **2022**, *11*, 1981. https://doi.org/10.3390/electronics11131981

Academic Editor: João Soares

Received: 29 May 2022
Accepted: 22 June 2022
Published: 24 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

China has built the world's largest electronic toll collection (ETC) system for expressways and deployed more than 20,000 gantry devices on 160,000 km of expressways nationwide. There are more than 200 million ETC OBU devices, with average daily ETC transaction data of nearly 1 billion [1]. The ETC transaction data record almost all vehicles' traffic conditions on expressways and can be used for expressway traffic flow prediction [2,3], transit time estimation [4,5], traffic demand visualization [6], etc. The data are expected to provide important information services for intelligent driving on expressways.

However, due to equipment failure, traffic jams, wireless crosstalk and other reasons, there are inevitably missed transactions and false transactions in the expressway ETC system. Obviously, when the ETC system is used to provide decision-making information for intelligent driving on expressways, the error rate and missing rate of ETC data need to be strictly controlled, and the error and missing ETC data must depend on the context of trajectory semantic structure to be accurately detected. As the key problem of time-series data processing, trajectory outlier detection is a popular topic at home and abroad, among which the dynamic time warping (DTW) algorithm [7] has been widely studied and applied as one of the mainstream methods of trajectory similarity measurement. Other such methods include user similarity analysis [8,9], travel mode analysis [10,11], and travel

route recommendation [12]. However, due to the huge amount of ETC transaction data and the lack of systematic data feature modeling, when some provincial platforms tried to use mainstream algorithms such as DTW to measure the error rate and missing rate of province-wide expressway ETC transaction data, a series of problems such as long time and low efficiency was encountered. Therefore, at present, the quality of ETC transaction data for national expressways is basically unknown, which seriously affects the efficiency of the ETC system of expressways and restricts the realization of the application of ETC transaction data.

In this work, aiming at the demand of intelligent driving expressway decision-making information based on ETC transaction data, we deeply studied the characteristics of ETC transaction data and designed a SegrDTW algorithm for segmented similarity recognition according to the characteristics of the ETC transaction data, through which we could quickly detect the problems of missed transactions and false transactions in the ETC transaction data. For the characteristics of the ETC transaction data, we used end-to-end bidirectional matching to extract dynamic step length and construct DTW search space windows, which reduced the computing time of the traditional DTW algorithm. By identifying the characteristics of the missed transactions and false transactions, the abnormal data in ETC transaction data culd be detected quickly. The algorithm proposed in this paper makes it possible to measure the abnormal events in expressway ETC transaction data and solves the key technical problems of the large-scale commercial application of expressway ETC transaction data.

The main contributions of this work are as follows:

- By improving the DTW algorithm, we proposed a SegrDTW algorithm that effectively reduces the time complexity of DTW.
- We also proposed an ETC data governance method that can effectively identify the false transactions and missed transactions in ETC data.

The rest of the article is arranged as follows. In the second part, the characteristics of the ETC transaction data are studied, the related definitions are given, and the ETC abnormal transaction data detection modeling is established. The third part is the design and analysis of the SegrDTW algorithm. In the fourth part, the experimental verification and result analysis are carried out using expressway field data from Fujian Province. In the final part, a summary of the full text is given, and the future work is prospected.

2. Modeling Abnormal ETC Transaction Data Detection

ETC transaction data include a large amount of vehicle traffic information and toll information, which can be used to reconstruct the time a vehicle passed through the toll stations and gantry along the way through the fields of vehicle identity document (ID), toll station equipment ID, gantry equipment ID, transaction time, etc. It is then possible to estimate the real-time position and speed of the vehicle and the traffic running situation of each section of the expressway. It can provide important decision-making information services for intelligent driving, such as reminding that there are turtle-speed vehicles in front and speeding vehicles approaching behind. According to the service demand of the intelligent driving decision-making information service, in order to systematically study the characteristics of ETC transaction data, this paper gives the following definitions.

Definition 1 (ETC transaction data). *According to the service demand of intelligent driving assistance decision information, expressway ETC transaction data* $(EData)$ *mainly includes three fields:* VID, SID *and Time, which are vehicle ID, toll station or gantry ID and transaction time.*

According to the VID field, the vehicle model, media access control address (MAC) and other information from the vehicle-mounted onboard unit (OBU) device in $EData$ can be further determined. Similarly, according to the SID field, we can further determine the longitude LNG and latitude LAT of an ETC toll station or gantry. Among them, we regard the toll station as a special gantry.

Definition 2 (the section). *The road section between two adjacent gantries on expressways is defined as section QD*:

$$QD = \langle SID_1, SID_2 \rangle \tag{1}$$

As shown in Figure 1, at present, the coverage distance of a road section usually ranges from several kilometers to more than 10 kilometers. According to the demands of the expressway intelligent driving assistant decision information service, the gantry position deployment can be increased or adjusted in the later stage.

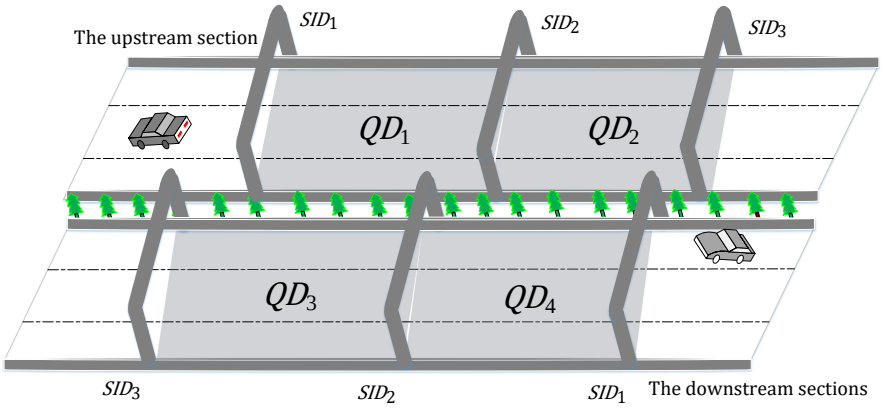


Figure 1. A schematic diagram of the sections.

The expressway ETC system inevitably has missed and false transactions; for example, when a vehicle passes through SID_2 on the upstream road, there can be a missed transaction due to the shelter of the big truck, or there can be a false transaction caused by SID_2 near the downstream road. Obviously, we cannot detect missed transactions and false transactions from a single point of data, so we need to rely on the semantic structure of the trajectory to achieve detection. Therefore, taking the toll station as the starting point and the end point, we give the definitions of transaction trajectory and OD path as follows.

Definition 3 (the transaction trajectory). *The EData sequence of the OD path formed by the vehicle VID from the entrance toll station SID_o to the exit toll station SID_d is called the transaction trajectory, ETraj*:

$$ETraj = \langle EData_o,\ EData_1,\ \ldots,\ EData_n,\ EData_d \rangle \tag{2}$$

where $EData_o$ and $EData_d$ are transaction data generated by SID_o and SID_d, respectively, and $EData_1$ to $EData_n$ are transaction data generated by several gantry frames between OD paths.

Obviously, all vehicles need to enter and exit the expressway from the toll station, and generally, it can be considered that there will be missed transactions at the toll station. Therefore, the *EData* is OD constrained; that is, the head and tail nodes of the transaction trajectory *ETraj* are all toll booth frames, and there will be no missed transactions or missed transactions.

Definition 4 (OD path). *The main topological path formed by all the entrance toll station data from the starting station SID_o to the ending station SID_d is called OD path ODLJ*:

$$ODLJ = \langle SID_o, SID_1, \ldots, SID_n, SID_d \rangle \tag{3}$$

In Figure 2, the red stations indicate the toll stations from beginning to end in an OD path, and the green stations indicate the passing entrance stations. At the same time, we call all the OD paths on the expressway OD path sets.

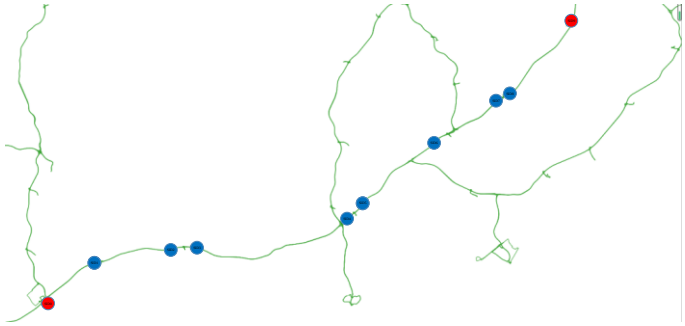


Figure 2. Topology diagram of an OD expressway route.

According to the above definition, the problem of detecting missed and false transaction data in *EData* can be transformed into a problem of comparing transaction trajectory and OD path; that is, all ETC data are normalized into a transaction trajectory set, and then abnormal trajectories can be detected by comparing that set with the OD path set. Therefore, how to improve the efficiency of abnormal trajectory detection becomes the key to identifying the of *EData* error and missing rates quickly and effectively.

One of the research directions of anomaly detection is realizing anomaly detection by measuring trajectory similarities in time series data. Mainstream algorithms include dynamic time warping (DTW), edit distance on real sequence (EDR), Hausdorff distance (Hausdorff), etc. [13]. On the basis of the algorithm, Mao et al. [14] proposed a Markov decision process-based detection method for abnormal spatial trajectories of road networks. The method classifies all the collected vehicle trajectory data according to the time sequence and then divides the trajectory data according to different vehicle trajectories to detect abnormal values. However, this method is not suitable for road network conditions, which change dynamically and are not suited to modeling dynamic incremental anomaly detection. George et al. [15] proposed an improved iterative DTW method according to the abnormal signal trajectory of layered ore body modeling to solve the problem of the large differences between the two signal lengths that greatly improved the accuracy of abnormal trajectory matching. However, this method also has some shortcomings in application, such as how to solve the sparseness problem when it is applied to traffic trajectory data. Ghersi et al. [16] proposed an automatic method of gait cycle extraction and analysis based on a DTW algorithm. This method combined the previous strategy with the dynamic time warping function, supplemented the segmentation method based on gait events, and tested it on undamaged and damaged people in order to overcome the need to calibrate or rely on the predefined threshold and reduce the error of step detection. However, this method does not consider the detection of turning stride or non-horizontal plane. Kumar et al. [17] proposed a DTW-based method, trajDTW, that used Dijkstra algorithm to calculate the similarities in road network constraint trajectories. This method is suitable for a large number of overlapping trajectories in dense road networks, supports the analysis of passengers' movement patterns and the planning of better bus routes, and also provides references for real-time route predictions for taxi passengers. When Hollerbach et al. [18] studied the single-valued arrangement of ion trajectories, they used the DTW algorithm to correct the changes in the sharing differences between lossless ion separations. In this method, single-valued alignment (SVA) and nonlinear dynamic time warping (DTW) data processing methods are used to correct the changes between IMS separations. However, the effect of this method for separation in a wide mobility range is much worse. Ruble et al. [19] proposed an improved DTW algorithm combined with hyper-dimensional Bayesian time mapping for mapping spatio-temporal trajectory signals with inserted or deleted elements. This method calculates the common underlying signal between two sequences, and it is superior to the DTW algorithm in sequence mapping and classification. However, when the signal-to-noise ratio of unconstrained sequences is low, the performance of this algorithm is poor. Cabanas et al. [20] proposed a DTW-based backtracking method to solve the problem

that the audio symbols in audio trajectories cannot be synchronized in music scores. This method adds a trajectory stage with a configurable delay. The small increment of delay will improve accuracy to a certain extent, but the algorithm has a delay and cannot be used in sensitive applications. As DTW is widely used in trajectory matching, a series of improved algorithms [21–25] was proposed by Emmon Keogh at the University of California for improving the retrieval efficiency of the DTW algorithm or the reliability of detection. Their team optimized the performance of the DTW algorithm by many methods. They improved the speed and accuracy of DTW optimization by indexing global and local data structures, parallel computing, time series clustering, and introducing parameters as a search mechanism for distance measurement. In addition to the mainstream DTW algorithm, other algorithms are also applied to the field of abnormal trajectory detection. Liu et al. [26] used the Hausdorff algorithm to establish the matching degree of the amplitude characteristics of the transient components of zero-sequence current trace at both ends of the feeder area of a distribution network. Through matching analysis, the identification ability of the fault sections in different environments was improved. This method adopts the comprehensive evaluation analysis method of wave channel model and ideal value approximation, which provides a reference for airlines to solve the problem of the comprehensive evaluation of flying quality. However, the application of the wave channel model is limited, and the accuracy of the evaluation results needs to be improved. Teng et al. [27] integrates a K-means thought and Hausdorff algorithm, constructs data segments using flight trajectory data, calculates the scores of multiple flight parameters in the data segment of level flight according to Brin channel theory, and finally determines the weights of each parameter through an "entropy weight method" to obtain index data reflecting the level flight quality of different UAV controllers. However, the robustness of the algorithm needs to be improved due to abnormal interference. In view of the fact that the current qualitative evaluation methods are time-consuming and laborious and that fair and objective evaluation is difficult, Qian et al. [28] selected the evaluation index by analyzing the data demand of pilot handling quality evaluation. The improved G1 method is used to select the weights, and the evaluation index system of pilot handling quality is established, so as to obtain an intuitive evaluation system. However, the accuracy of the algorithm evaluation is greatly affected by the selected indicators, and when the indicators are not rigorous enough, the evaluation results are not convincing. Wang et al. [29] proposed an ordered generalized Hausdorff distance for measuring the geometric quality of point cloud data and proposed a model representation of time series according to the changing trend. The method can quickly calculate the distance of model and has no multi-resolution characteristics, which can effectively identify the abnormal degree of data. However, because the choice of resolution is very important for obtaining the similarities between sequences, the model distance of this method can only represent the difference between two sequences with the same length; it cannot reflect the differences between sequences with different lengths. Qiu et al. [30] proposed the OBF-EDR magnetic anomaly signal similarity measurement method based on the combination of orthogonal basis function (OBF) decomposition and EDR. In this method, the discrete basis function coefficients are obtained by the orthogonal basis function decomposition of magnetic anomaly signals, and the signal-to-noise ratio of the discrete basis function coefficients is improved according to the characteristic that background noise is independent of the basis function. The edit distance method is used to calculate the similarities in the discrete basis function coefficients and thus indirectly realize the similarity measurement of magnetic anomaly signals. Cao et al. [31] proposed an aircraft trajectory anomaly detection method based on the combination of a multidimensional outlier descriptor (MOD) and bidirectional long-short memory network (Bi-LSTM). Firstly, the trajectory deviation detection is transformed into the trajectory density classification problem, and a multidimensional anomaly descriptor is designed for the trajectory deviation detection model so as to effectively detect the trajectory deviation and trajectory outliers. Among them, the distance between time series is evaluated using the DTW similarity function instead of the traditional Euclidean distance, which solves the problem that

Euclidean distance is difficult to accurately calculate. However, the corresponding time of the model is long, and it is difficult to guarantee the real-time detection. Liu et al. [32] proposed a new asynchronous trajectory-matching method (PTSCTM) based on piece-wise spatio-temporal constraints, which is used for AIS trajectory reconstruction and ship abnormal behavior discovery. Based on the spatio-temporal constraints, this method finds the optimal trajectory matching point in the candidate matching point set with time and space distance and uses it to calculate the similarities in trajectory matching. In order to reduce the time to find matching points after segmentation, the similarity between trajectory segments only depends on the matching degree of trajectory points in trajectory segments. When the trajectory length is short, the calculation results will deviate.

However, $EData$ has the characteristics of large scale and high density, being composed of discrete points obtained by sampling with discrete and continuous mathematical characteristics. Obviously, mining abnormal data directly from a large number of transactions is a heavy workload and is inefficient. When processing unique $EData$ to meet the demand of massive real-time ETC abnormal data detection, the sequential algorithm of spatio-temporal deformation trajectory detection needs to be further optimized and improved so as to quickly and accurately identify abnormal data and missed/false transactions.

The mainstream algorithms, DTW, Hausdorff, EDR, etc., that are used in the field of spatio-temporal deformation detection have a time complexity of $O(n^2)$, so it takes a great deal of time to process big trajectory data. In order to effectively improve the efficiency and accuracy of the abnormal trajectory detection of $EData$, a restricted search space algorithm, SegrDTW, based on the $EData$ of expressways is proposed. It takes SID_o and SID_d as starting points, carries out the end-to-end bidirectional matching to extract the dynamic step size, cuts the matching range of DTW by constructing a segmented matching feature window, and makes an accurate spatial search for the mismatched trajectory sequence, thus greatly reducing unnecessary matrix calculation and time complexity.

SegrDTW not only verifies the similarities between the path sequence $ODLJ$ and the vehicle trajectory sequence $ETraj$ but also identifies abnormal situations in the different parts of the trajectory, such as missed transactions or false transactions. For the degree of abnormal deviation, we made the following definitions.

Definition 5 (abnormal degree). *When the VID of the vehicle passes through an ODLJ path, the corresponding transaction data ETraj will be generated, and the proportion of the missed or false transaction data in the whole transaction data is called the abnormal degree. The abnormal degree of ODLJ can be expressed as follows:*

$$e(ODLJ) = \sum_{i=1}^{n} e(QD)/n \tag{4}$$

where $e(QD)$ is the abnormal degree of a single section in$ODLJ$. $e(QD) \in [0, 1]$ means that the value of $e(QD)$ is between 0 and 1 when there is a missed transaction or a false transaction in a section QD.

Through the above definitions, it can be concluded that $EData$ has the following main characteristics: (1) $EData$ is spatio-temporal sequence data. (2) As vehicles must pass through toll stations, they are constrained by OD characteristics. (3) There are many trajectories on the OD path, but the main OD path has the largest number of trajectories. (4) Due to the missed and false transactions between the ETC vehicle-mounted equipment and the roadside equipment, OD path and ETC gantry topology only have semantic coupling.

Therefore, faced with large-scale and dense $EData$, it is necessary to find a way to adapt and process the data with these characteristics and at the same time reduce the workload of algorithm processing so as to obtain the corresponding results more directly and conveniently.

3. An Algorithm for Detecting Abnormal Data of ETC

The general framework of the abnormal ETC transaction data detection algorithm is shown in Figure 3. According to the characteristics of *EData*, this paper proposes an improved algorithm, SegrDTW, based on DTW. The algorithm matches the similarities of vehicle trajectories passing on *ODLJ* path and quickly calculates the unmatched and dissimilar trajectory segments between *ETraj* and *ODLJ* through the method of a limited constraint feature window. According to the characteristics of *EData*, an abnormal trajectory identification algorithm is designed to identify the missed transactions/false transactions in *EData*. The final steps are to map and label these abnormal *EData* and verify the missed/false transactions to confirm the accuracy of the results.

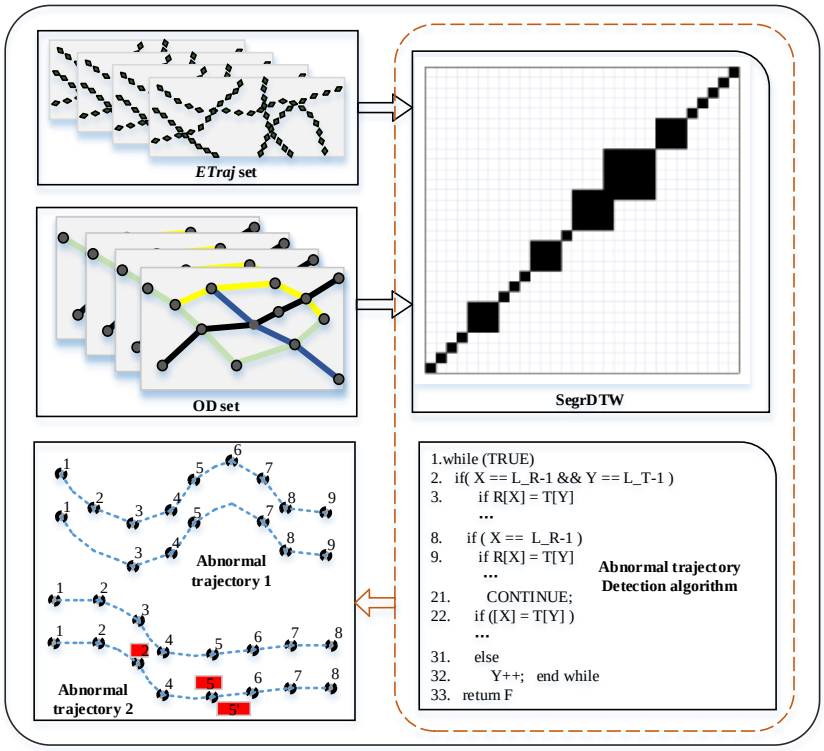


Figure 3. The framework diagram of the abnormal ETC transaction data detection algorithm.

3.1. Trajectory Similarity Matching Algorithm

First, an improved SegrDTW algorithm is proposed to detect the anomalies in ETC transaction data. By matching the similarities between *ODLJ* trajectory and *ETraj* vehicle trajectory, we can identify the abnormal trajectories in *ETraj* vehicle trajectory. We define the similarities in the trajectories as follows.

Definition 6 (Trajectory similarity). *There are overlapping or abnormal trajectories between ODLJ trajectory and ETraj vehicle trajectory, and DTW distance can be used to quantify the similarity between vehicle trajectory and frame trajectory:*

$$Sim(ODLJ, ETraj) = DTW(ODLJ, ETraj) = D(n, m) \tag{5}$$

where D is the regular matrix composed of path sequence ODLJ and the transaction trajectory ETraj , and n and m are the sizes of the matrix, i.e., the lengths of the two sequences, respectively. Using the dynamic programming algorithm, we can find a shortest path SP from the matrix, which is defined as follows:

$$D(i,j) = \begin{cases} Dist\left(ODLJ_i - Etraj_j\right) + min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\}, if\ i \neq 0\ and\ j \neq 0 \\ 0, if\ i = j = 0 \\ \infty, if\ i = 0\ or\ j = 0 \end{cases} \quad (6)$$

where Dist is the Euclidean distance between two data points; i,j are denoted as the i-th data in the ODLJ and the j-th data in the ETraj trajectory sequence, respectively. The pathSP must satisfy 3 major constraints:

Boundaries: SP must start at $P_1 = (1,1)$ *and end at* $P_{max} = (n,m)$.
Monotonicity: $P_k^{row} - P_k^{row} \geq 0$ *and* $P_k^{col} - P_{k-1}^{col} \geq 0$, *ensuring chronological matching and avoiding cross-matching.*
Continuity: $P_k^{row} - P_k^{row} \leq 1$ *and* $P_k^{col} - P_{k-1}^{col} \leq 1$, *ensuring adjacent point matching and avoiding cross-point matching.*

The DTW algorithm is a classic spatio-temporal sequence alignment template-matching algorithm. It requires building a distance matrix from the two trajectories and finding the path with the smallest sum of the element values from the upper left corner to the lower right corner. The DTW algorithm requires path planning for each iteration, which may lead to problems such as a high calculation complexity and a lot of space. In the expressway ETC system, the ETC transaction sequence cannot use the traditional one-to-one match method to measure the similarities due to the existence of missing transactions and false transactions in the ETC transaction data. The DTW algorithm can lengthen or shorten the sequence axis to process the matching operations in spatiotemporal sequences of different lengths. However, the traditional DTW algorithm adopts point-by-point matching technology and uses dynamic programming technology to complete the optimal path search. For example, for a vehicle trajectory *ETraj* with a length of M and an OD path $ODLJ$ with a length of N, the initial subsequence size of DTW is $N - M + 1$. The computational effort to do the work of matching known sequences would be $O(N \times M)$. Therefore, the optimization strategy for DTW mainly focuses on two strategies: early discard (lower bound) and limited search space.

Based on the DTW algorithm and the characteristics of ETC transaction data, this work compares *ETraj* with $ODLJ$ to find the abnormal state of transactions in the *ETraj* and then designs a restricted search space strategy. Due to the existence of missed transactions or false transactions, which are occasional and uncertain, the trajectory lengths of *ETraj* and OD path sequence $ODLJ$ are not necessarily equal. When there are abnormal data in the vehicle trajectory sequence *ETraj*, there must be a situation where the unmatched trajectory is inconsistent with the expected corresponding $ODLJ$. To solve this problem, we propose a concept of dynamic search matching step length R, which can reduce unnecessary matching loss by limiting the range of dynamic matching step length R. In the next step, the algorithm first identifies the matching points P in *ETraj* trajectory points without leakage and error within the range R and takes these points P as the matching feature points of vehicle trajectory *ETraj* and OD path sequence $ODLJ$. Finally, the search window W is generated according to the matched feature points, and the recognition range of missed and false transactions is limited to W, which reduces the amount of calculation and improve the efficiency of the algorithm.

Definition 7 (the dynamic search step size). *There are cases of missed transactions and false transactions in the vehicle trajectory sequence ETraj, which leads to inconsistencies between the ETraj and the ODLJ. In order to solve the problem of unequal length comparison, this paper puts forward the concept of dynamic search step size, that is, the comparison starting point of two trajectories plus the absolute value of the difference between two trajectories, which is called the dynamic search step size R of transaction gantry comparison. R is defined as follows:*

$$R = |Len(ODLJ) - Len(ETraj)| + 1 \quad (7)$$

where Len is a function of obtaining the length of the trajectory sequence.

By adding the absolute length deviation between the $ODLJ$ and the $ETraj$, and the boundary length, the range value of the dynamic search step R is obtained. The dynamic search step R is used to limit the maximum search length of the mismatch trajectory between the $ETraj$ and the $ODLJ$. When the $ETraj$ is longer than the $ODLJ$, there must be some false transaction events in the $ETraj$. Therefore, when matching the trajectory point $ODLJ_i$ traversing in $ODLJ$ with the trajectory point of $ETraj$, we can start from the i-th trajectory point $ODLJ_i$ and add the length of the dynamic search step R as the search range, so as to quickly determine the matching feature points in the range. Similarly, when the length of the $ETraj$ is shorter than the length of the path $ODLJ$, it can be inferred that there must be some missed transaction events in the vehicle trajectory $ETraj$. Additionally, when we search for the trajectory point $ODLJ_i$ that traverses $ODLJ$ with the trajectory point of $ETraj$, since there are missed transaction events (missed data) in $ETraj$, the current i-th trajectory point $ODLJ_i$ needs to be compared with some subsequent data in the vehicle trajectory $ETraj$, so we limit the matching range of the vehicle trajectory sequence $ETraj$ by the length of the dynamic search step R, which can speed up the search of matching feature points that match $ODLJ_i$. If the matching is successful within the range of R, this point is marked as the feature point of successful matching, and if the matching is unsuccessful within the range of R, this boundary is marked as a failure matching. After completing the retrieval of $ODLJ_i$, the next step is to retrieve the next trajectory point $ODLJ_{i+1}$.

The steps of matching point retrieval and recognition are described below. Since the trajectory $ETraj$ formed by $EData$ has a starting point and an ending point, and the spatio-temporal data of the trajectory satisfy the definition of the midpoint in Euclidean space, we define the matching feature point P in the trajectory sequence $ETraj$ and $ODLJ$, where the Euclidean distance between $ETraj$ and $ODLJ$ is 0, that is, $P|dist(ETraj, SODLJ) = 0$. However, the sequence lengths of $ETraj$ and $ODLJ$ are different. In addition to the feature matching points, there are some abnormal points in the transaction record sequence that can not match $ODLJ$. By comparing these points in European distance space, it is impossible to confirm the cause of the anomaly; that is, it is impossible to know which gantry the vehicle passed through to cause the anomaly. In order to solve this problem effectively, we adopted the two-way traversal search strategy of dynamic step R tolerance range. As shown in the following Figure 4, $ETraj$ and $ODLJ$ use the Euclidean distance calculation functions $dist(ETraj, ODLJ_1)$ and $dist(ETraj, ODLJ_2)$, respectively, taking $ODLJ_1$ as the starting point and $ODLJ_2$ as the ending point to calculate the matching feature points. When $dist(ETraj, ODLJ_1) \neq 0$ or $dist(ETraj, ODLJ_2) \neq 0$, that is, the matching of traversing trajectory points is unsuccessful, the algorithm tries to continue to verify the matched feature points within the limited range, dynamic step R, and at the same time mark and identify the intermediate abnormal data.

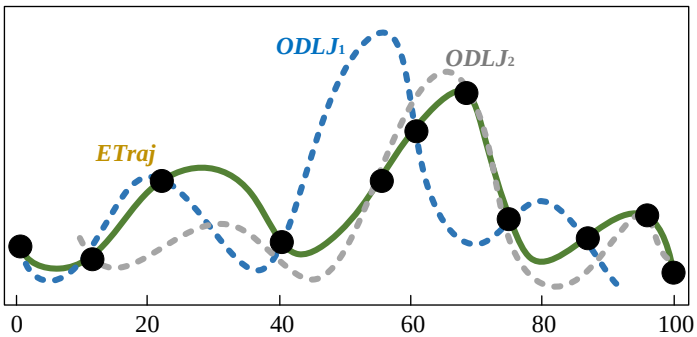


Figure 4. The segmentation of the matching feature window.

After identifying the feature points (i.e., the trajectory points with no abnormalities), these matched feature points are used as the boundaries of the anchor points, and the

window is constructed for the area where the distance between these anchor points is greater than 1 by algorithm. These windows are used to restrict the operation areas to identify abnormal data such as missed transactions and false transactions.

Algorithm 1 for realizing matching feature points P and establishing search window W of abnormal trajectory range is as follows.

Algorithm 1. Algorithm for establishing anomalous trajectory range retrieval window

Input: $ETraj$ (transaction trajectory), $ODLJ$ (gantry sequence)
Output: W (window of abnormal trajectory range.)
Parameters:

1: I (Initialize gantry sequence pointer)
2: J (Initialize vehicle trajectory pointer)
3: R (dynamic search step size and the value is ABS ($ETraj$. length $ODLJ$. length) + 1
4: FA (feature value storage array)
5: EA (auxiliary mark storage array)
6: //Mark the characteristic points matching the rack from beginning to end.
7: **FOR** (I = 0, J = 0; I < $ODLJ$. length && J < $ETraj$. length; I++, J++)
8: **IF**(gantry trajectory point $ODLJ[I]$ is matched with vehicle trajectory point $ETraj[J]$)
9: Save the location of $ODLJ[I]$ in FA and the location of $ETraj[J]$ in EA
10: **ELSE**
11: **WHILE**(the index value of the vehicle trajectory point is still in the range of R)
12: If the gantry trajectory point $ODLJ[I]$ matches the current vehicle trajectory point, record the position of the gantry trajectory point and the vehicle trajectory point, otherwise, try to locate the next trajectory point to continue matching.
13: **END WHILE**
14: **END FOR**
15: //Matching feature points from back to front.
16: **FOR** (I = $ODLJ$.length −1 J = $ETraj$.length −1; I > −1 && J > −1 I−−, J−−)
17: Proceed from step 7 to step 14 from back to front.
18: **END FOR**
19: //Building a window of abnormal features.
20: **FOR** (I = 1; I < FA.length; I++)
21: **WHILE** (J < EA.length)
22: Get the driving trajectory sub-sequence and a gantry frame sub-sequence.
23: W.ADD (AFC (trajectory sub-sequence, gantry frame sub-sequence))
24: **END WHILE**
25: **END FOR**
26: **RETURN** W

From step 7 to step 18 of the algorithm, we first traverse in the positive and negative directions and use the Euclidean distance ratio to determine the matching feature points. If the match fails, we will confirm whether there are matching points within the range of step size R. As shown in Figure 5, $\langle d_0, \ldots, d_n \rangle$ and $\langle q_0, \ldots, q_m \rangle$ are the trajectory points of $ETraj$ and $ODLJ$, respectively. The Euclidean distance algorithm is used to calculate the matching trajectory point with $p = 0$ in the range of R (the orange position point in Figure 5). If there is no matching point in the constraint range of R (the green or blue search area in Figure 5), mark (d_i, q_j) and search the trajectory point of (d_{i+1}, q_{j+1}) to find the matching feature points. The algorithm performs forwards and backwards traversal to ensure that the intermediate data without corresponding relationship can be identified in the matching process and that all the segmented matching points with Euclidean distance equal to 0 in the trajectories of $ETraj$ and $ODLJ$ are obtained. From step 20 to step 25, the algorithm constructs a window W according to the identified segmentation matching points as boundaries:

$$W_n = \sum_{i,j=1}^{row,col} min\{D(i-1,j), D(i,j-1), D(i,j)\} \quad (8)$$

where D is the grid area formed by the driving trajectory sub-sequence of $ODLJ$ and the sub-sequence of $ETraj$.

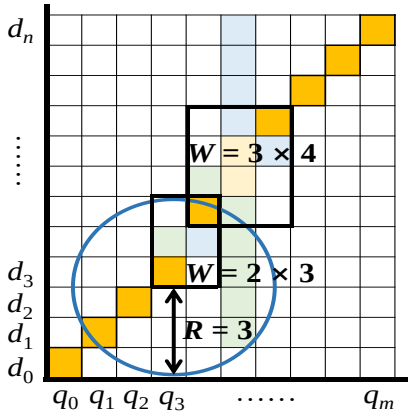


Figure 5. Schematic diagram of finding and confirming matching points.

By constructing the minimum window W, $W = r_{row} \times r_{col}$, $r_{row} = \langle r_1, r_2, r_3, \ldots, r_m \rangle$, $r_{col} = \langle r_1, r_2, r_3, \ldots, r_n \rangle$. Using the abnormal feature identification algorithm AFC (step 23), the abnormal data in the window are detected by path planning to identify the missed or false transaction data. Step 26 returns the window W of the recognized data feature result.

3.2. Performance Analysis of SegrDTW Algorithm

The feature window is constructed as the region between (d_2, q_2) and (d_6, q_4) and the region between (d_7, q_5) and the feature matching points (d_{10}, q_7), as shown in Figure 6. After the window is built, the similarity of dynamic path planning can be calculated by the DTW algorithm. The window built by $ODLJ$ and $ETraj$ can be divided into two categories. One is the matching window ξ, that is, Euclidean distance $dist(d, q) = 0$, with time complexity $O(1)$, such as $\xi_1 = dist(d_1, q_1)$, $\xi_2 = dist(d_2, q_2)$, and $\xi_3 = dist(d_6, q_4)$, and the other is the sequence difference window η, which has the retrieval distance $dist(d, q) > 0$ with two divided matching windows, such as $\eta_1 = dist([d_3, d_4, d_5], q_3)$ or $\eta_2 = dist([d_8, d_9], q_6)$.

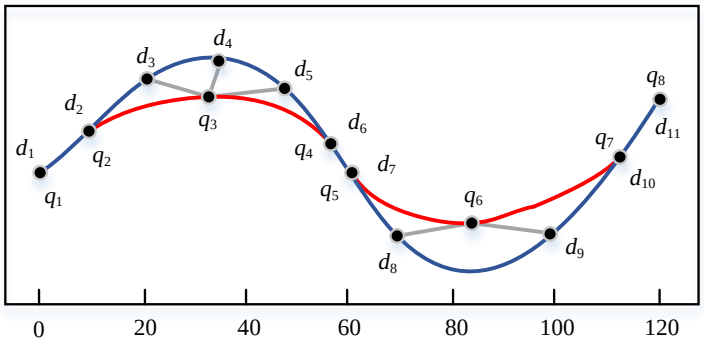


Figure 6. A projection comparison between two trajectory sequences in the window.

The similarity of the two trajectories is calculated as follows:

$$argminD = \sum_{i=1,j=1}^{m,n} D_{i,j} \tag{9}$$

where $D_{i,j} = \langle \xi_1, \ldots, \xi_m, \eta_1, \ldots, \eta_n \rangle$.

In order to prove that the retrieval efficiency of the proposed SegrDTW algorithm is better than the DTW algorithm, the following related theorems are given and theoretically proved in this paper.

Theorem 1. *Given a standard trajectory ODLJ=$\langle q_1, q_2, q_3, \ldots, q_m \rangle$ of sequence length m and an ETC trajectory ETraj = $\langle \xi_1, \xi_2, \ldots, \xi_i, \eta_{i+1}, \eta_{i+2}, \ldots, \eta_n \rangle$ of sequence length n, the necessary time complexity is $O_{SegrDTW}(ODLJ, ETraj) < O_{DTW}(ODLJ, ETraj)$.*

Proof of Theorem 1. There are normal or abnormal trajectory points in the *ETraj* trajectory sequence. Let ξ be the normal trajectory points in the section and $P(\lambda_1)$ be the probability of ξ appearing in the section, that is, $\xi \sim P(\lambda_1)$. η represents the abnormal trajectory points in the cross section, and $P(\lambda_2)$ is the probability of η appears in the section, that is, $\eta \sim P(\lambda_2)$. If $P(\lambda_1)$ and $P(\lambda_2)$ are independent of each other, then there is $\xi + \eta \sim P(\lambda_1 + \lambda_2)$. ξ , η is a discrete random variable, and all possible k values of $\xi + \eta$ are 1, 2, ... ; event $\{\xi + \eta = k\}$ is the sum of mutually exclusive events $\{\xi = i, \ \eta = k - i\}$, $(0 \le i \le k)$. Because ξ and η are independent of each other:

$$\begin{aligned} P(\xi + \eta = k) &= \sum_{i=1}^{k} P(\xi = i, \eta = k - i) \\ &= \sum_{i=1}^{k} P(\xi = i) P(\eta = k - i) \\ &= \sum_{i=1}^{k} \frac{\lambda_1^i e^{-\lambda_1}}{i!} \frac{\lambda_2^{k-i} e^{-\lambda_2}}{(k-i)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{i=1}^{k} C_k^i \lambda_1^i \lambda_2^{k-i} \\ &= \frac{(\lambda_1+\lambda_2)^k}{k!} e^{-(\lambda_1+\lambda_2)} \end{aligned} \tag{10}$$

It can be deduced that $\xi + \eta \sim P(\lambda_1 + \lambda_2)$ and the event conform to the Poisson distribution. Assuming $\xi \sim P(\lambda_1)$, the similar paths are calculated by Euclidean distance, the distance is 0, and the time complexity is $O(1)$. Let $\eta \sim P(\lambda_2)$ calculate similar paths through the DTW algorithm. When $r_{row} < m$ and $r_{col} < n$, there must be $O(r_{row}, r_{col}) < O(m, n)$. Because $\xi \sim P(\lambda_1)$ and $\eta \sim P(\lambda_2)$ are independent events, the trajectory generation is not affected, so the similarity calculation of ξ can be neglected, and it takes a great deal of time to search and calculate only by considering the sequence of η in the trajectory. Therefore, it can be concluded that the SegrDTW algorithm is superior to the DTW algorithm in execution efficiency. □

Algorithm complexity analysis: For the ETraj sequence of length M and the $ODLJ$ sequence of length N, the traditional DTW algorithm uses point-by-point matching technology and dynamic programming technology to complete the optimal path search to find the matching path, and its time complexity is $O(N \times M)$. In order to improve the search efficiency, in the SegrDTW calculation process, $ETraj$ and $ODLJ$ are first calculated to match the trajectory points in the forward and reverse directions. These two steps are based on the path length N of $ODLJ$, and the total computational complexity is $2 \times O(N)$. Then, according to these matching trajectory points, there are n windows $W(r \times r)$ with missed transactions or false transactions in the section, and the number of normal section windows $W(1 \times 1)$ is $(N - n)$. At last, only n windows $W(r \times r)$ need to be calculated to match the path. Therefore, the total time complexity of the SegrDTW algorithm can be calculated as follows:

$$\begin{aligned} T &= 2N + n \times O(W_{r \times r}) + (N - n) + n \times O(r^2) \\ &= 2N + (N - n) + n \times O(r^2) + n \times O(r^2) \\ &= 3N + n \times [2O(r^2) - 1] \end{aligned} \tag{11}$$

It can be seen from Equation (11) that the total time complexity of the SegrDTW algorithm depends on the path length N of $ODLJ$, the number n of recognized windows and the dynamic step length r of windows. Through the statistical analysis of the trajectory data for Fujian Province in the early stage, the statistical data of the incidence of abnormal trajectory windows and the dynamic step length r of built windows are shown in Table 1.

Table 1. The dynamic search matching range length and abnormal data rate of a three-day ETC transaction trajectory in Fujian Province.

Date	Number of N	Lower Limit of Dynamic Search Matching Range R	Upper Limit of Dynamic Search Matching Range R	Trajectory Data Anomaly Rate Statistics
3 September 2020	41,860	3	5	4.73%
4 September 2020	40,651	3	4	4.65%
5 September 2020	41,769	3	5	5.30%

According to the statistics of the existing ETC trajectory data, the value range of r is in [3,5]. Compared with the path length N of $ODLJ$ reaching more than 40,000, the value of r is much smaller than N, which means its influence is almost negligible; that is, the time complexity of $O(r^2)$ is about equal to $O(1)$. Because the ratio of abnormal window n in the total trajectory matching is 4–6% of N, and n is much smaller than N, that is, $n << N$. Therefore, the above conditions can be substituted into Formula 11 $\approx 3N + n \times O(1) \approx N$, and the time complexity of the SegrDTW algorithm can be deduced as $O(N)$.

3.3. Algorithm for Identifying Abnormal Transaction Data

After matching the abnormal trajectories, we can obtain the trajectory sub-sequence data with missed transactions or false transactions in the trajectory sequence $ETraj$ of each vehicle in the trajectory set. Next, we need to identify causes that may lead to these missed and false transactions. $EData$ has spatio-temporal attributes (transaction time and latitude and longitude position of the gantry), and the space distance of the gantry deployed in the adjacent lanes of the expressway is much smaller than that of the same lane. Based on these spatio-temporal features, the abnormal trajectory sequence data can be identified, i.e., the abnormal transaction data (missed transaction or false transaction) identification algorithm will compare and verify the spatio-temporal semantics of the constructed abnormal window W and the trajectory point data in the trajectory sequence $ODLJ$. The identification formula for verifying abnormal characteristic value F is as follows:

$$F = \begin{cases} 1, & ODLJ_{w[i]} = ETraj_{w[j]} \\ 0, & otherwise \end{cases} \tag{12}$$

Through the positive and negative directions of two-way verification, the characteristic F of missed transactions or false transactions is obtained, and the data on these missed transactions or false transactions are tagged and returned to the result set. The abnormal transaction data feature recognition algorithm is as follows Algorithm 2.

According to the matching distance set of the vehicle trajectory sequence $ETraj$ and the gantry sequence $ODLJ$, the algorithm identifies the abnormal features of the vehicle trajectory. According to the principle of continuity and monotonicity in the DTW algorithm, from step 5 to step 9, through the comparison of distances in three directions and according to the minimum distance, the running trajectory of the vehicle is calculated, and the spatio-temporal information of the SID passing through the gantry is compared with the trajectory point of $ETraj$. Through the comparison, the flow status of the transaction data in this SID is obtained, and 0/1 is marked, where 1 is no exception and 0 is a missed transaction. For example, 01/10/11 indicates that false transactions exist in this SID, and other statuses indicate the detection of other errors. From step 12 to step 18 of the algorithm, the window boundary is processed. When calculating the route, if the gantry sequence set reaches the last value while the vehicle trajectory set does not reach the last value, only the rest of the vehicle trajectory set needs to be circulated; otherwise, only the rest of the gantry sequence set needs to be circulated. At last, the algorithm returns the feature sequence.

Algorithm 2. Feature recognition algorithm for abnormal transaction data

Input: *ODLJ* (gantry sequence set)
ETraj (vehicle trajectory set)
Output: *F* (Identification label results)

1: Initialize the gantry sequence pointer *I*
2: Initialize the vehicle trajectory pointer *J*
3: **WHILE**(TRUE)
4: //Matching the front window trajectory
5: **IF** (the current trajectory point match the current value of the entry sequence)
6: the position of the current trajectory point is recognized as the normal trajectory point and stored in the result set F.
7: **ELSE**
8: Match and verify the spatio-temporal attribute of the trajectory point pointed by *J* with the value on the right side of the gantry sequence pointed by *I* and the value near the top of the window, confirm whether the trajectory is a missed transaction or a false transaction through the spatial position, and output the result of identifying the result set *F*.
9: **END IF**
10: The gantry pointer *I* and the vehicle trajectory pointer *J* point to the next set of trajectory data.
11: //Returns the output results of the window boundaries in three cases
12: **IF** (the gantry pointer *I* and the vehicle trajectory pointer *J* both reach the end point)
13: Comparing the last object of the vehicle trajectory *ETraj* with the last object of the gantry sequence *ODLJ*, and outputting the result to F.
14: **ELSE IF** (the sequence pointer I reaches the end point)
15: Compare the last object of *ETraj* with the remaining object of *ODLJ*, and output the result to the recognition result set.
16: **ELSE IF** (the vehicle trajectory pointer *J* has reached the end point)
17: Comparing the last object of the gantry sequence *ODLJ* with the remaining vehicle trajectory *ETraj*, and outputting the result to *F*.
18: **END IF**
19: **END WHILE**
20: **RETURN** *F*

4. Results

The conditions of this experiment are a Windows 10 system in a Java language development environment with an Intel Core 2.6 GHz processor and 16 GB memory on a Lenovo computer.

4.1. Data Sources and Experimental Settings

The data set contains about 60,344,500 pieces of real transaction data collected by Fujian Expressway Information Technology Company through a ETC gantry system from 3–12 September 2020, which is converted into 9,057,200 transaction trajectories, and the gantry topology information of Fujian Province. Among these data, the original transaction data table contains 103 fields, recording the information of each part of the vehicle and gantry, including license plate number, gantry id, transaction data, gantry time, latitude and longitude, etc. (see Table 2).

Table 2. ETC gantry system transaction data attributes.

Attribute Name	Examples	Attribute Name	Examples
Trade ID	340×98	OBU Plate	Blue Fujian A1×45
Trade time	5 September 2020 21:29:26	Vehicle Class	1
Flag ID	25××14	Enter Time	5 September 2020 20:23:51
Flag Type	0	Enter Station	25×7
Flag Index	1	OBU ID	13B×××D6
LAT	118.56××	LNG	24.85×××

4.2. Experiment and Discussion on Verification of SegrDTW Model

4.2.1. Introduction of the Experimental Data

Before verifying our algorithm experiment, the data anomalies of 9,057,200 trading trajectories from 3–12 September 2020 are analyzed by DTW and other algorithms. The data used in the experiment account for 27% of the total data, as shown in Figure 7a. In addition, it is confirmed that there is a correlation between the vehicle journey and the number of gates passing through in the ETC transaction data used. As shown in Figure 7b, when a vehicle passes through more ETC gantries, the error rate of the vehicle's trajectory will also increase.

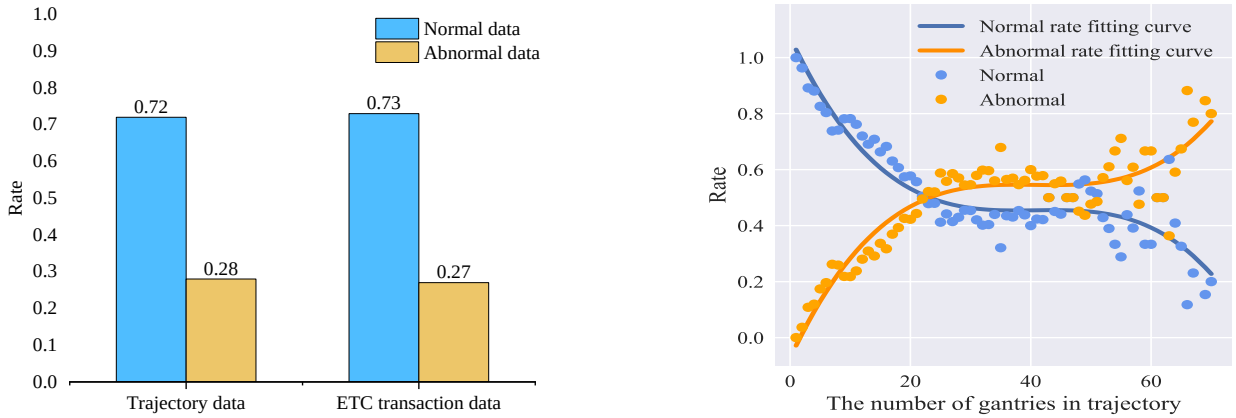


Figure 7. Analysis of the experimental data: (**a**) overview of abnormal transaction data; (**b**) the correlations of the numbers of ETC gantries in the vehicle trajectories.

4.2.2. Comparison of the Algorithms' Performance

In order to verify the efficiency of SegrDTW algorithm in detecting and identifying missed/false transaction data, this experiment selected four sections in Fujian Province: Yiban-Haicang hub, Gaochang-Dapu hub, Xianyoubangtou-Fuzhounan, and Nanpingbei-Minhouganzhe as test sections. The data covers Fuzhou, Xiamen, Quanzhou, Sanming, Putian and Nanping.

This experiment verifies the performance of the algorithm proposed in this paper from three aspects: the retrieval efficiency of trajectory data with specified OD path length, the retrieval efficiency of long-distance trajectory data and the retrieval efficiency of large amounts of data.

The first part is the verification of the retrieval efficiency of specified OD path trajectory. The time efficiency of SegrDTW is compared with that of the DTW, Hausdorff and EDR algorithms, and the results are shown in Table 3. Compared with the other algorithms, SegrDTW has obvious advantages in efficiency. Experiments can prove the effectiveness of the SegrDTW algorithm model in the application of spatial-temporal data similarity, which ensures the robustness of retrieval and effectively improves the running efficiency. For example, on the Yiban-Haicang hub, the retrieval time for SegrDTW is 6.9, 7.3 and 10 times shorter than the times for DTW, Hausdorff and EDR, respectively.

Table 3. The algorithm retrieval times.

Sections	SegrDTW	DTW	Hausdorff	EDR
Yiban-Haicang hub	0.18 s	1.25 s	1.31 s	1.80 s
Mawei-Ningdexi	0.24 s	1.15 s	1.38 s	1.78 s
Xianyou Bangtou-FuZhounan	0.15 s	0.77 s	0.98 s	1.28 s
Nanpingbei-Minhou ganzhe	0.21 s	1.14 s	1.21 s	2.25 s

From a set of one-day transaction data for Fujian Province, a total of 16,528 pairs of ODs can be obtained. According to the average retrieval time of OD pairs in the above four sections, the abnormal data detection time of all OD pairs in the whole province can be estimated. The EDR algorithm takes 29,378.52 s (about 8.16 h), the Hausdorff algorithm takes 20,164.16 s (about 5.6 h), the DTW algorithm takes 17,808.92 s (about 4.95 h), and SegrDTW takes 3222.96 s (about 0.86 h). The SegrDTW algorithm proposed in this paper saves much time in detecting the daily transaction data for the whole province. If the ETC gantry transaction data anomaly detection is applied in the whole country, the retrieval efficiency of the method proposed in this paper will have greater advantages.

We also verify the performance of the algorithm according to the long-path transaction data (long data retrieval performance). In the experiment, the OD sequence of gantry in the Fujian section of Shen-Hai Expressway was selected, with the lengths of 15, 25, 35, 45 and 55 gantries, to match and detect the vehicle trajectories. The results are shown in Figure 8. The horizontal axis represents the length of the sequence, while the vertical axis represents the algorithm running time. As can be seen from the figure, compared with other algorithms, the running time of the SegrDTW algorithm has a linear relationship with the length of the sequence, and the range of change is very small, which is close to a linear relationship and has little impact. This is mainly because our algorithm adopts the dynamic search step size search, which can reduce the amount of calculation to the greatest extent.

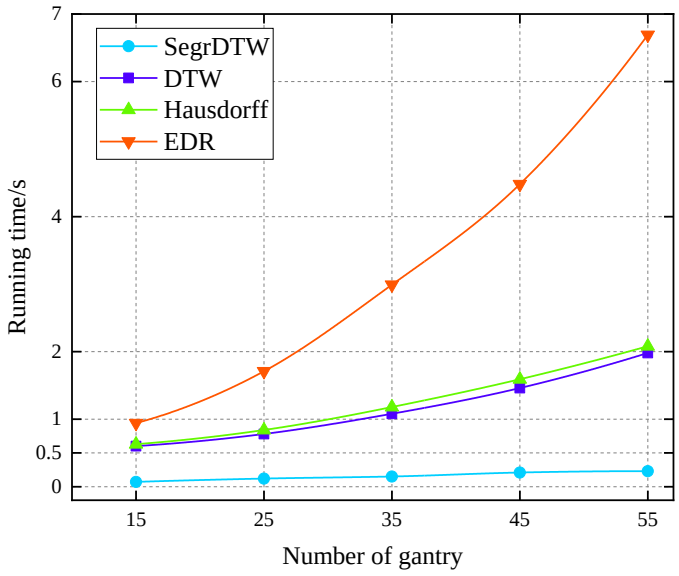


Figure 8. Number of gantries tested for run time.

In view of the abnormal feature detection requirements of the ETC transaction data, we also set up a large number of data to verify the processing pressure of the algorithm. Specifically, the data for Quanzhou to Xiamen from September 3 to September 12 were selected for the experiment. The results are shown in Figure 9. The horizontal axis represents the data volume of the 10-day trajectory, and the vertical axis represents the running time. It can be seen from the figure that the running times for the EDR, DTW and Hausdorff algorithms increase exponentially with increasing amounts of trajectory data, which indicates that the amount of data has a great influence on the running time. There is a linear relationship between the running time of the SegrDTW algorithm and the amount of data. As the increase of data volume, the range of change has increased slightly and tends to be stable. This algorithm has strong performance and stable operation. This is because the SegrDTW algorithm greatly increases the proportion of discarding and reducing unnecessary calculations, thus showing the greater compactness of the lower bound. When the spatio-temporal complexity is low, the dissimilar sub-sequences are discarded as early as

possible, which reduces the subsequent DTW calculation, so that the retrieval time will not obviously increase with increases in the number.

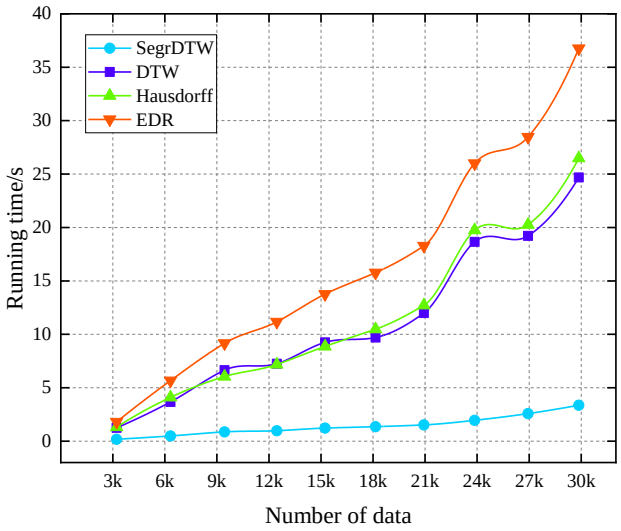
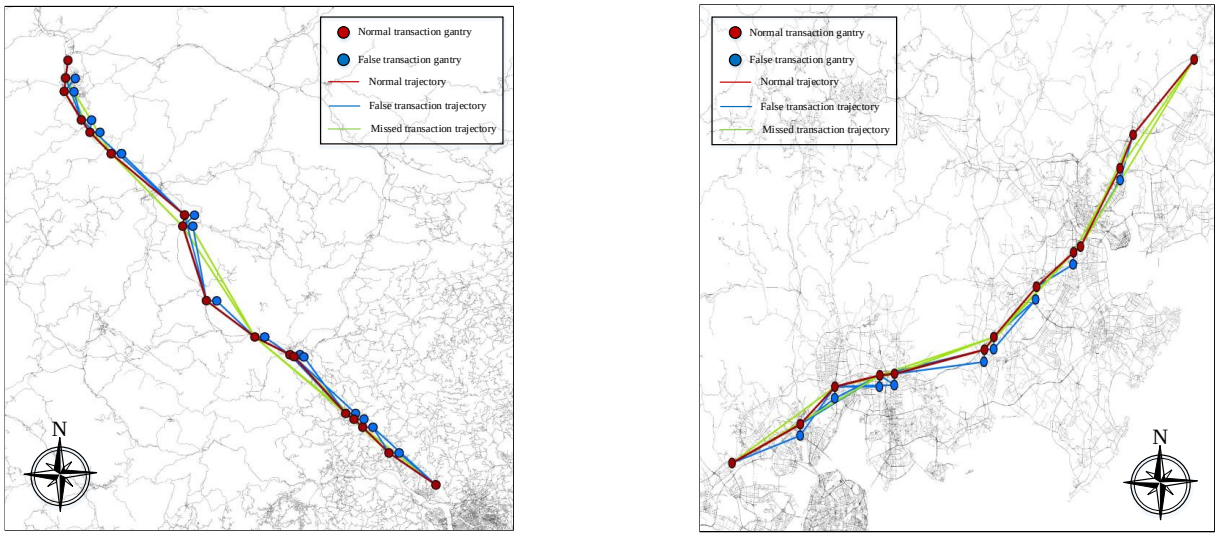


Figure 9. Running efficiency tests on different data volumes.

4.2.3. Identify the Abnormal Transaction Trajectory

In this experiment, according to the transaction data of *ETraj*, we also tested the accuracy of the identification of missed transactions and false transactions. In order to verify the recognition accuracy and the effectiveness of the algorithm for abnormal data, four representative road sections were selected in the experiment: sugarcane section from Nanpingbei-Minhou Ganzhe, Yiban to Haicang hub section, Mawei-Ningdexi section and Xianyou Bangtou-Fuzhounan section. The abnormal detection projections of missed transactions and false transactions caused by vehicles passing through these road sections are shown in Figure 10.



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

Figure 10. *Cont.*

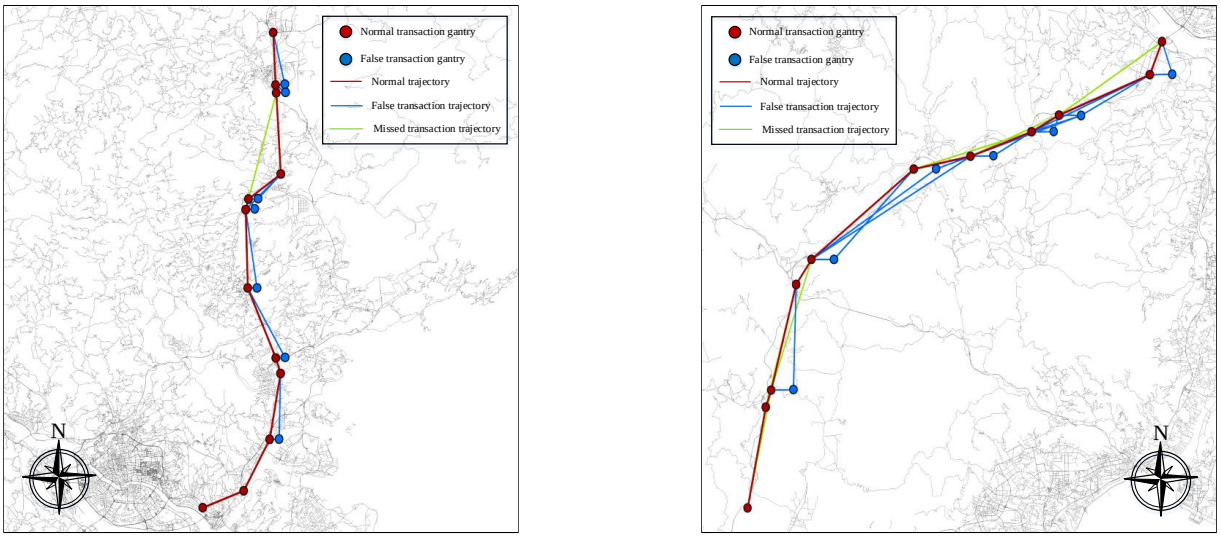


Figure 10. The abnormal transaction trajectory detection findings: (**a**) Nanpingbei-Minhouganzhe; (**b**) Yiban-Haicang. (**c**) Mawei-Ningdexi; (**d**) Xianyoubangtou-Fuzhounan.

The figures show the driving trajectories of all vehicles passing through the selected four OD paths over three days. The red dot is the ETC gantry on the normal road, while the blue dot is the gantry on the adjacent road. The red trajectory line is the trajectory of the vehicle passing through the correct gantry (overlapping with the *ODLj*), while the blue and green trajectories are the trajectories of the vehicles with false transactions or missed transactions. The abnormal transaction data account for the total number of trip transactions as shown in Figure 11.

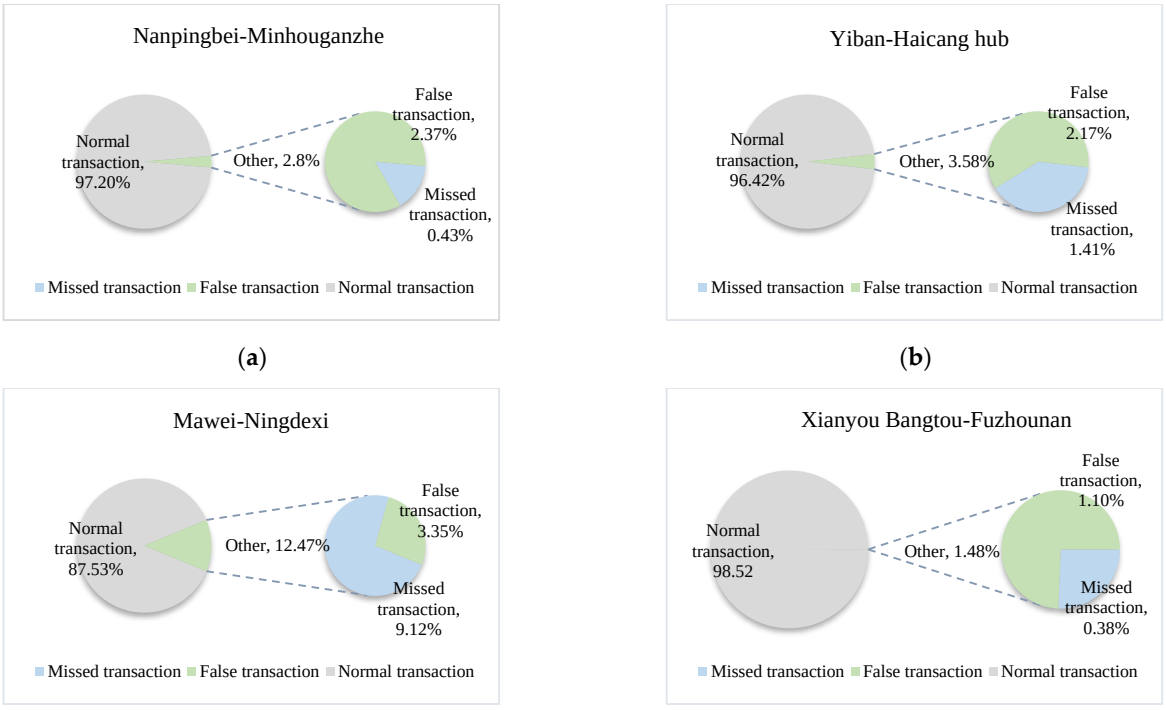


Figure 11. Abnormal transaction ratios of the driving trajectories. (**a**) The abnormal proportions of missed transactions and false transactions in Nanpingbei-Minhou Ganzhe are 0.43% and 2.37%; (**b**) The

abnormal proportions of missed transactions and false transactions in Yi-ban-Haicang hub are 1.41% and 2.17%; (**c**) The abnormal proportions of missed transactions and false transactions in Mawei-Ningdexi are 9.12% and 3.35%; (**d**) The abnormal proportions of missed transactions and false transactions in Xianyou Bangtou-Fuzhounan are 0.38% and 1.10%.

EData is detected and analyzed using abnormal transaction data detected using the SegrDTW algorithm, and the abnormal degree $e(ODLJ)$ of vehicle $ETraj$ transaction data is between 4.65% and 5.30%. The identified abnormal data can be further cleaned or repaired to provide reliable data for the follow-up intelligent expressway driving decision information service.

4.2.4. Discussion of the ETC Abnormal Transaction Data Detection Algorithm

In order to further verify the effectiveness of the SegrDTW algorithm, 3168 normal trajectory data and 3168 abnormal trajectory data are selected in the experiment. Compared with the accuracy of the DTW algorithm, the experimental results are shown in Table 4. It can be seen from Table 4 that the normal detection accuracy is 100%, and the abnormal trajectory detection accuracy is 99.46%. The detection accuracy of the DTW algorithm is the same as that of SegrDTW, which indicates that SegrDTW can achieve fast detection and ensure high retrieval accuracy.

Table 4. Accuracy statistics.

Class	Number	DTW					SegrDTW				
		TP	FP	TN	FN	Accuracy	TP	FP	TN	FN	Accuracy
Normal	3168	0	0	3168	0	100%	0	0	3168	0	100%
Abnormal	3168	3156	12	0	0	99.46%	3156	12	0	0	99.46%

We use the original feature data and the abnormal feature data set generated to evaluate the accuracy of SegrDTW. The statistical results for the false or missed transaction data and the correct transaction data are shown in Table 5, with a total of 6336 items. Table 5 shows that the detection accuracy rates of false transactions and missed transactions for the DTW and SegrDTW algorithms are both above 99%, which means both algorithms can effectively detect abnormal transaction data. At the same time, the recall rates of false transactions and missed transactions are both lower than 100%, indicating that there will be a small number of samples detected as missed transactions in the false transaction detection test, and a small number of samples in the missed transactions will also be detected as false transactions. The DTW algorithm uses a global search method with precision of 100%, while SegrDTW uses a dynamic radius search method, and the search at the boundary will produce wrong detection, resulting in a lower accuracy rate than DTW. The F1-score results show that the difference between the two algorithms is less than 0.02, which proves that the robustness of the SegrDTW model is slightly lower than that of the DTW algorithm.

Table 5. The detection results for abnormal data types.

Abnormal Class	DTW				SegrDTW			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
False transaction	0.9996	1	0.9916	0.9958	0.9906	0.9995	0.9833	0.9913
Missed transaction	0.9998	1	0.9932	0.9966	0.9914	0.9996	0.9848	0.9921

5. Conclusions

In this paper, we propose an SegrDTW algorithm. The dynamic step size is used to limit the window of similarity retrieval, which effectively reduces the time complexity of the DTW algorithm and can be used for the rapid detection of abnormal ETC data. We evaluated our model on a large-scale traffic dataset. The experiment results show that our

proposed method significantly outperforms several competing methods. Although this algorithm improves the efficiency of abnormal data retrieval, it also has some shortcomings. For instance, the algorithm adopts dynamic radius segmentation retrieval, and there is still a wrong detection at the boundary of segmented trajectory sequence. In the future, we will focus on improving the accuracy of the boundary recognition of the abnormal trajectory range retrieval window and the applicability of the algorithm.

Author Contributions: Conceptualization, F.Z. and F.G.; methodology, F.G. and S.L.; software, F.G.; validation, L.L., S.L. and X.Y.; formal analysis, S.L. and C.Z.; investigation, F.G. and J.W.; resources, F.Z.; data curation, F.G.; writing—original draft preparation, F.G. and S.L.; writing—review and editing, F.Z., F.G. and S.L.; visualization, L.L. supervision, J.W.; project administration, F.G. and X.Y.; funding acquisition, F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the National Natural Science Foundation of China (41971340), the Special Funds for the Central Government to Guide Local Scientific and Technological Development (2020L3014), the 2020 Fujian Province "Belt and Road" Technology Innovation Platform (2020D002), and the Provincial Candidates for the Hundred, Thousand and Ten Thousand Talent of Fujian (GY-Z19113). Crosswise project (No. GY-H-21021).

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Fujian Expressway Information Technology Co., Ltd. and are available from the authors with the permission of Fujian Expressway Information Technology Co., Ltd.

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

References

1. Qian, M. Analysis of multi-dimensional data fusion and application of ETC portal system. *China ITS J.* **2021**, *6*, 109–112.
2. Zhao, Y.Y.; Chen, Y.J.; Guan, W. Prediction Model of ETC Short Term Traffic Flow Based on Multidimensional Time Series. *J. Transp. Syst. Eng. Inf. Technol.* **2016**, *16*, 191–198.
3. Chen, Z.; Wu, B.; Li, B.; Ruan, H. Expressway Exit Traffic Flow Prediction for ETC and MTC Charging System Based on Entry Traffic Flows and LSTM Model. *IEEE Access* **2021**, *9*, 54613–54624. [CrossRef]
4. Chiou, J.M.; Liou, H.T.; Chen, W.H. Modeling Time-Varying Variability and Reliability of Freeway Travel Time Using Functional Principal Component Analysis. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 257–266. [CrossRef]
5. Chen, L.W.; Chen, D.E. Exploring spatiotemporal mobilities of highway traffic flows for precise travel time estimation and prediction based on electronic toll collection data. *Veh. Commun.* **2021**, *30*, 100356. [CrossRef]
6. Tsung, C.K.; Yang, C.T.; Yang, S.W. Visualizing potential transportation demand from ETC log analysis using ELK stack. *IEEE Internet Things J.* **2020**, *7*, 6623–6633. [CrossRef]
7. Müller, M. *Dynamic time warping. Information Retrieval for Music and Motion*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 69–84.
8. Lv, M.; Chen, L.; Chen, G. Mining user similarity based on routine activities. *Inf. Sci.* **2013**, *236*, 17–32. [CrossRef]
9. Xiao, X.; Zheng, Y.; Luo, Q.; Xie, X. Inferring social ties between users with human location history. *J. Ambient. Intell. Humaniz. Comput.* **2014**, *5*, 3–19. [CrossRef]
10. Liu, Y.; Kang, C.; Gao, S.; Xiao, Y.; Tian, Y. Understanding intra-urban trip patterns from taxi trajectory data. *J. Geogr. Syst.* **2012**, *14*, 463–483. [CrossRef]
11. Fang, Z.; Shaw, S.L.; Tu, W.; Li, Q.; Li, Y. Spatiotemporal analysis of critical transportation links based on time geographic concepts: A case study of critical bridges in Wuhan, China. *J. Transp. Geogr.* **2012**, *23*, 44–59. [CrossRef]
12. Zheng, Y.; Xie, X. Learning travel recommendations from user-generated GPS traces. *ACM Trans. Intell. Syst. Technol. (TIST)* **2011**, *2*, 1–29. [CrossRef]
13. Li, H.L.; Liang, Y.; Wang, S.C. Review on dynamic time warping in time series data mining. *Control Decis.* **2018**, *33*, 1345–1353.
14. Mao, J.Y.; Wu, H.; Sun, W.W. Vehicle trajectory anomaly detection in road network via Markov decision process. *Chin. J. Comput.* **2018**, *41*, 1928–1942.
15. George, M.A.; Silversides, K.L.; Zigman, J.; Melkumyan, A. Bedding Angle Identification from BIF Marker Shales via Modified Dynamic Time Warping. *Math. Geosci.* **2021**, *53*, 1567–1585. [CrossRef]
16. Ghersi, I.; Ferrando, M.H.; Fliger, C.G.; Arenas, C.F.C.; Molina, D.J.E.; Miralles, M.T. Gait-cycle segmentation method based on lower-trunk acceleration signals and dynamic time warping. *Med. Eng. Phys.* **2020**, *82*, 70–77. [CrossRef]
17. Kumar, D.; Wu, H.; Rajasegarar, S.; Leckie, C.; Krishnaswamy, S.; Palaniswami, M. Fast and scalable big data trajectory clustering for understanding urban mobility. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 3709–3722. [CrossRef]
18. Hollerbach, A.L.; Conant, C.R.; Nagy, G.; Monroe, M.E.; Gupta, K.; Donor, M.; Giberson, C.M.; Garimella, S.V.B.; Smith, R.D.; Ibrahim, Y.M. Dynamic Time-Warping Correction for Shifts in Ultrahigh Resolving Power Ion Mobility Spectrometry and Structures for Lossless Ion Manipulations. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 996–1007. [CrossRef] [PubMed]

19. Ruble, M.; Hayes, C.E.; Welborn, M.; Zajic, A.; Prvulovic, M.; Pitruzzello, A.M.; Hayes, E. Hyperdimensional bayesian time mapping (hyperbat): A probabilistic approach to time series mapping of non-identical sequences. *IEEE Trans. Signal Process.* **2019**, *67*, 3719–3731. [CrossRef]
20. Cabanas-Molero, P.; Cortina-Parajón, R.; Combarro, E.F.; Alonso, P.; Bris-Peñalver, F.J. HReMAS: Hybrid real-time musical alignment system. *J. Supercomput.* **2019**, *75*, 1001–1013. [CrossRef]
21. Keogh, E.; Wei, L.; Xi, X.; Vlachos, M.; Lee, S.H.; Protopapas, P. Supporting exact indexing of arbitrarily rotated shapes and periodic time series under euclidean and warping distance measures. *VLDB J.* **2009**, *18*, 611–630. [CrossRef]
22. Lin, J.; Keogh, E.; Lonardi, S.; Lankford, J.P.; Nystrom, D.M. Visually mining and monitoring massive time series. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; pp. 460–469.
23. Bagnall, A.; Lines, J.; Bostrom, A.; Large, J.; Keogh, E. The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Min. Knowl. Discov.* **2017**, *31*, 606–660. [CrossRef]
24. Zakaria, J.; Mueen, A.; Keogh, E. Clustering Time Series Using Unsupervised-Shapelets. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM), Brussels, Belgium, 10–13 December 2012.
25. Hao, Y.; Campana, B.; Keogh, E. Monitoring and Mining Animal Sounds in Visual Space. *J. Insect Behav.* **2013**, *26*, 466–493. [CrossRef]
26. Liu, L.; Zhang, T.; Ru, B. A flying qualities assessment model based on multiparameter integration. *Comput. Eng. Sci.* **2016**, *38*, 1262–1268.
27. Teng, H.L.; Li, B.W.; Gao, Y.; Yang, D.; Zhang, Y. Quality evaluation model of unmanned aerial vehicle's horizontal flight maneuver based on flight data. *J. Beijing Univ. Aeronaut. Astronaut.* **2019**, *45*, 2108–2114.
28. Qian, X.; Cai, Z.C. A Mathematical Model for Evaluating Pilot Controlling Quality of Military Aircraft Pilot Controlling. *Ordnance Ind. Autom.* **2014**, *33*, 16–18.
29. Wang, D.; Rong, G. Pattern distance of time series. *J. Zhejiang Univ. Eng. Sci.* **2004**, *38*, 795–798.
30. Qiu, J.; Ou, J.D.; Xie, D.; Wang, Z.; Du, J. A Similarity Measurement Method for Magnetic Anomaly Signal Under Low Signal-to-noise Based on Orthogonal Basis Function–Edit Distance. *J. Electron. Inf. Technol.* **2022**, *44*, 745–753.
31. Cao, Y.; Cao, J.; Zhou, Z.; Liu, Z. Aircraft Track Anomaly Detection Based on MOD-Bi-LSTM. *Electronics* **2021**, *10*, 1007. [CrossRef]
32. Liu, C.; Wang, J.; Liu, A.; Cai, Y.; Ai, B. An Asynchronous Trajectory Matching Method Based on Piecewise Space-Time Constraints. *IEEE Access* **2020**, *8*, 224712–224728. [CrossRef]