Journal of Traffic and Transportation Engineering(English Edition)
交通运输工程学报(英文版)
ISSN 2095-7564,CN 61-1494/U

《Journal of Traffic and Transportation Engineering(English Edition)》网络首发论文

题目：　Explore the lane change trajectory pattern driven by full-time domain trajectory data

作者：　Min Zhang，Yuhan Nie，Bo Wang，Chi Zhang，Yuming Zhou，Yijing Zhao

网络首发日期：　2024-09-18

引用格式：　Min Zhang，Yuhan Nie，Bo Wang，Chi Zhang，Yuming Zhou，Yijing Zhao．Explore the lane change trajectory pattern driven by full-time domain trajectory data[J/OL]．Journal of Traffic and Transportation Engineering(English Edition)．https://link.cnki.net/urlid/61.1494.U.20240914.1413.005

Published online：2024-09-18 09:27:11
URL：https://link.cnki.net/urlid/61.1494.U.20240914.1413.005

Original research paper

Explore the lane change trajectory pattern driven by full-time domain trajectory data

Min Zhang[a] ,Yuhan Nie [a,d], Bo Wang[b,c,d] ,Chi Zhang [b,d,*] ,Yuming Zhou[b] , and Yijing Zhao[a,d]

[a] School of Transportation Engineering, Chang'an University, Xi'an, China;

[b] School of Highway, School of Highway, Chang'an University, Xi'an, China;

[c] School of Civil and Environmental Engineering, Nanyang Technological University, Singapore, Singapore;

[d] Engineering Research Center of Highway Infrastructure Digitalization, Ministry of Education, Xi'an, China

Highlights

- A universal method framework for identifying lane changing trajectory patterns on highways.
- Optimized the penalty term for change point detection by considering driving scenarios.
- Similarity matching takes into account the stages and fluctuation characteristics of lane changing.

Abstract

Due to the increasing density and complexity of the highway network, a deep understanding of the characteristics of lane changing (LC) behavior is crucial for road refinement design. The emergence of full time domain trajectory big data provides unprecedented opportunities for in-depth research on highway safety geometry design. This article proposes a method framework for extracting LC trajectory patterns to explore the combination trajectory patterns during the LC process. This article achieves subdivision in driving mode detection by using the Adaptive Pruned Exact Linear Time (APELT) algorithm to detect change points, taking into account the short sequence features of LC. In order to achieve the classification of segmented fragments, we report a clustering technique based on similarity matching (SM), which can effectively avoid the problem of excessive distortion in similarity measurement. The results indicate that APELT technology has certain advantages in F1 score and accuracy of LC pattern recognition, which is more in line with reality. The kappa score based on similarity matching is greater than 0.8, indicating high accuracy of pattern recognition. This study provides a novel data mining method for a comprehensive understanding of lane changing behavior under full time domain big data, which will provide reference for road design in complex scenes.

Keywords:

Geometric design; Road safety; Lane change behavior; Change point detection; Similarity measurement.

*Corresponding author.
E-mail addresses: minzhang@chd.edu.cn (M. Zhang), 19907429968@163.com (Y. Nie), wb1010110wb@chd.edu.cn(B. Wang), zhangchi@chd.edu.cn (C. Zhang), 17502968661@163.com(Y. Zhou), 2022034010@chd.edu.cn(Y. Zhao)

1 Introduction

The geometric design of roads affects the safety of lane changing (LC) behavior, as different geometric designs result in different frequency distributions of LC trajectory patterns (De Almeida, Pimentel Vasconcelos, and Cesar Bastos Silva 2018). The LC trajectory pattern is a specific driving behavior that occurs repeatedly by one or more drivers (Tselentis and Papadimitriou 2023). Previous studies have shown that the patterns of LC behavior are related to driver behavior, aggression, gear prediction, and drowsiness (Wang and Wu 2023; Wu and Chang 2022), and the fundamental reason for the frequent occurrence of risky LC trajectory patterns is the instability of vehicle driving caused by geometric design. For example, the LC trajectory pattern in the diversion area of interchanges may be similar to the LC pattern in emergency stop events on straight sections, while the distribution of LC patterns in interchanges A and B exhibits heterogeneity. Therefore, identifying LC trajectory patterns in a full time domain environment and investigating the frequency and distribution of different patterns can explore the geometric design of roads or the differences in LC behavior in specific scenarios (different types of work areas and interchanges), thereby guiding the refined design of highway scenes.

Pattern recognition of LC trajectory belongs to a branch of driving pattern recognition (Tselentis and Papadimitriou 2023). Previous research on LC pattern recognition includes identifying driving operation patterns and vehicle interaction patterns during LC, which is considered as a problem of time series segmentation and similarity clustering (Elspas et al. 2021). Driving operation pattern For example, Chen (Chen et al. 2022) used the clustering framework to explore the correlation pattern between vehicle speed change and driving manipulation during LC. The vehicle interaction pattern of LC is a correlation pattern to study the spatial position of LC vehicles and the time when the surrounding vehicles are about to collide, and its research is very rich. The classic methods are point clustering method (Chen et al. 2021) and sequence segmentation method (Zhang et al. 2023). As for the research results of interactive pattern recognition, it is found that a common LC pattern is radical LC. In addition to two types of LC patterns, it also includes the pattern of LC trajectory. For example, Fan (Fan et al. 2022) uses unsupervised hybrid method to identify abnormal behaviors in LC, including abnormal hesitation or radicalization of speed or acceleration. Cheng (Cheng et al. 2016) tried to use DTW to solve the problem of identifying driving patterns and distinguish between speeding and illegal overtaking patterns. However, the track pattern recognition is mainly based on the simulation data or the data of limited length road sections, ignoring the LC pattern recognition in a large range of time and space. Although previous studies have shown very ideal results in various pattern recognition, each track pattern found cannot understand their relationship with road risk and geometric design (Tselentis and Papadimitriou 2023). Our research object is the LC trajectory in a wide range of time and space. Examples of possible detected patterns are repeated trajectory risk sequences such as emergency braking, rapid acceleration and sharp turn in different stages of LC (before and after LC, offset stage). These are all important aspects, because they will enable us to have a deeper understanding of the geometric design features that affect the risk of LC.

Pattern recognition of LC trajectory is a methodological challenge (Ali, Zheng, et al. 2020). Unsupervised learning and data mining techniques were first used to discover driving trajectory patterns and develop labeling schemes (Ali, Bliemer, et al. 2020). For example, the point clustering method was first applied to identify patterns, and Chen (Chen et al. 2021) used LC risk index combined with k-means algorithm to classify LC behaviors and identify LC risk periods. Later, the discovery of driving pattern is to use segmented aggregation approximation. The classic methods include Chen (Chen et al. 2022) dividing the LC sequence into blocks by using a hidden semi-Markov model based on hierarchical Dirichlet process, and then clustering the segments by using a potential Dirichlet assignment (LDA) model to identify different maneuvering segments of drivers in the LC process. And Zhang (Zhang et al. 2023) decomposed the LC scene based on Hidden Markov Model and Gaussian Mixture Model (GMM-HMM), and then applied dynamic time warping (DTW) K-means clustering to aggregate the primitives into 13 LC interaction patterns to identify potential risk segments. It is worth noting that the above-mentioned method based on GMM assumes that the observation data of each state is composed of multiple Gaussian distributions. For each time point, GMM can calculate the probability distribution of the observation data in each state, and HMM can extend the probability distribution to the state transition probability matrix, the observation probability matrix and the initial state probability distribution after GMM modeling. However, the method based on GMM-HMM assumes that the data obey Gaussian mixture model and Markov chain, which is suitable for identifying long driving sequences, but it is inconsistent with the actual situation when identifying short LC sequences. It needs to estimate a large number of parameters, which is very computationally intensive and may have the risk of over-fitting, so its application in full-time domain trajectory data is limited. On the other hand, LC pattern is a specific driving behavior, which should be recognized at a more detailed level, that is, within a very short driving time (within a few seconds) (Tselentis and Papadimitriou 2023). And PELT (Pruned Exact

Linear Time) is a change point detection method that realizes linear time complexity through pruning operation, which is suitable for short sequence segmentation. Its core idea is to minimize the cost function (Haynes, Fearnhead, and Eckley 2017). By adjusting the penalty term, it can improve the robustness to noise and outliers without making specific assumptions about data distribution (Shi and Morris 2021). Therefore, we consider the change point detection technology and prove that it is more suitable for distinguishing LC states. PELT algorithm is very flexible. We improved the PELT algorithm (Haynes et al. 2017) to cut the segments of LC process. This is very suitable for the LC motion process based on trajectory variation characteristics in the process of cutting LC, and the multi-dimension of features is considered. In addition, similar clustering method has been proved to be effective in clustering LC fragments after cutting (Hamedi and Shad 2022), but DTW is usually distorted. Just as Fan (Fan et al. 2022), the LC process can be divided into different stages, namely, maintenance, change, arrival and adjustment, according to the relationship between the vehicle and the centerline and other characteristics, and the interaction pattern between different LC stages and speeds is complex. Therefore, we solve the problem of excessive distortion in dynamic time warping by considering the phase of LC based on similarity matching (SM) method and adding constraints. In a word, this paper aims to consider the existing problems and propose a pattern recognition method for LC trajectory data in full time domain, so as to provide a method reference for LC behavior pattern recognition in acceleration, braking and turning for highway road geometric design.

A. Change point detection technology

At present, change point detection technology is mainly used in the field of 5G signal detection, aiming at identifying the time stamp when reconfiguration occurs in the operator network. In the process of LC, in order to increase traffic safety, it is also necessary to identify the state changes of LC process, estimate the sudden change time, identify the time periods of different states, and test the changes. The change point can be defined as an index in the time series, at which the statistical characteristics suddenly change (Li, Huang, and Wen 2023). A typical method of change point detection is to optimize the cost function through a set of possible change point indexes to minimize it. Previous research shows that compared with binary segmentation and Bayesian online change point detection technology, the PELT change point detection method using information standard as penalty term in the 5G field has higher accuracy in segmentation. In 2021, Shi et al. (Shi and Morris 2021) applied PELT algorithm to detect potential dynamic restart change points, thus maximizing the joint probability of piecewise continuous potential dynamic representation. They suggested using marginal likelihood as the scoring function of

PELT, thus avoiding the punishment based on the complexity of the model. In 2022, Panek (Panek, Jablonski, and Wozniak 2022) applied the supervision method to the penalty setting of PELT algorithm. In addition, sliding window change point detection technology and wavelet transform also have many applications. However, they (Zheng and Washington 2012) are not well adapted to multi-scale change point detection and are very sensitive to the selection of window size. PELT algorithm is an adaptive method, which can detect the change points of different scales (Li et al. 2023). It is reported that some work has been published on the penalty problem, which optimizes the selection of penalty values, such as ED-PELT (Haynes et al. 2017) with the best penalty setting. However, appropriate parameterization, especially penalty setting, is also needed to adapt to the task set. In order to solve this problem, this paper proposes an adaptive pruned exact linear time (APELT). From the perspective of LC behavior, the improvement proposed in this study is a supplement to the scene consideration in LC trajectory pattern extraction. In a word, the application of change point detection technology in identifying LC time series segmentation is not enough, although they are of great significance, which will be further proved in this work.

B. Clustering method based on similarity

The segmented fragments are multidimensional sequences with different lengths. Previous studies show that similarity clustering method can be used to classify fragments (Di et al. 2022; Qu et al. 2022; Zhang and Shi 2021). For example, Hamedi (Hamedi, Shad, and Ziaee 2022) use DTW-FCM technology to measure the position, speed and surrounding vehicle status as LC similarity dimensions. Long (Long et al. 2022) uses four-dimensional dynamic time warping (4DTW) to measure the distance between two four-dimensional time series. Zhang (Zhang et al. 2023) applied (DTW) K-means clustering to decompose LC primitives. DTW allows a certain degree of deviation on the time line, and can measure different lengths of time (Chen and Gao 2020). However, the disadvantage of using classical DTW is heavy computational burden, especially when dealing with multivariate time series. In order to avoid the shortcomings of the above-mentioned time series similarity measurement methods, this paper improves a distance measurement method based on similarity matching (SM) (Chen and Gao 2020). On the one hand, after improvement, not only the fluctuation characteristics of time series can be extracted, but also the range limitation of matching can be considered. On the other hand, it can match the tracks of the same LC stage as much as possible.

In addition to the challenge of methods, it is also very important to use what kind of data to identify patterns. Most researches on driving pattern recognition use natural driving experimental data collected from OBD devices or smart phone sensors (Panichpapiboon and Leakkaw 2020) and driving simulators (Jokhio et al. 2023). Among many data acquisition technologies of driving characteristics, the second-level high-precision coordinate positioning data of full-time domain and whole road section obtained by heavy freight floating car has a series of remarkable advantages. Compared with the general data, the data has the characteristics of full-time domain, wide range, high precision and high frequency, which is conducive to comprehensively and accurately grasping the LC characteristics of various interchanges and different sections of heavy trucks in the day and night environment. In addition, the analysis of LC behavior of heavy trucks can represent the most unfavorable influence of LC behavior in various scenes of expressways, which will guide the fine design of existing expressway geometry. Previous studies mainly collected speed and positive acceleration and used them for driving analysis, followed by negative acceleration (braking), time stamp and GPS coordinates (Cheng et al. 2016). Among them, speed and acceleration indicators are very important in personal driving risk safety assessment (Tselentis and Papadimitriou 2023). In addition, not all papers provide data collection frequency, but most papers use 1Hz collection frequency. This shows that 1Hz is an acceptable frequency, which balances the collection of noise data and insufficient information. In this paper, the four-dimensional characteristics of speed, lateral and longitudinal acceleration and heading angle change are considered to identify patterns. The data covers many scenes of road LC, including road air humidity, interchange, work area and so on. We also consider the difference between free LC and mandatory LC (Ali et al. 2021).

In order to better extract the trajectory pattern of LC process, this paper proposes a two-stage unsupervised learning framework (APELT-SM) based on change point detection and similarity matching. Under this framework, the first step is to use APELT technology to detect the change point and segment the LC multidimensional sequence. The second step is to apply SM method to cluster LC fragments. The application of this framework is expected to better identify the LC trajectory pattern under the background of full-time domain big data. The main contributions of this study include:

1) The change point detection technology is introduced into the LC sequence segmentation method, and an Adaptive Pruned Exact Linear Time technology suitable for LC sequence segmentation is proposed. The results of F1 and precision show that this algorithm is superior to other algorithms in identifying LC state, which is more in line with the actual situation.

2) Using SM-based method to cluster the segmented segments effectively solves the phenomenon of excessive distortion in distance measurement. The calculated similarity matrix is evaluated by Kappa, and the result is greater than 0.8, which shows that the measurement method of predefined LC trajectory similarity is reliable. Compared with the commonly used LC pattern extraction methods, our framework has the highest kappa score in LC duration, which shows that it has good performance in discovering LC trajectory patterns.

3) A LC data preparation method for full-time domain trajectory data is proposed. Our framework has been applied to analyze the influence of external environment and road geometry type on the frequency of LC pattern occurrence, which shows that the identified pattern is interpretable, which further enriches the application of high-frequency GPS data. The remaining parts of this article are as follows: Section 2 introduces our proposed method framework and provides a detailed introduction to the algorithms used in each stage. Section 3 introduces the preparation method for floating vehicle LC data and the evaluation criteria for the method. Section 4 introduces the experiment and results, while Section 5 discusses the experimental results. Finally, provide concluding comments in Section 6 of the Conclusion.

2. METHODOLOGY

The framework APELT-SM proposed in this paper is shown in Fig 1. Firstly, LC events are extracted from full-time domain trajectory data. Then, APELT technology is used to train the model and segment the LC trajectory, and the segmented segments are clustered based on SM matching similarity clustering method. Finally, we analyze the LC trajectory pattern of expressway in multiple scenes.

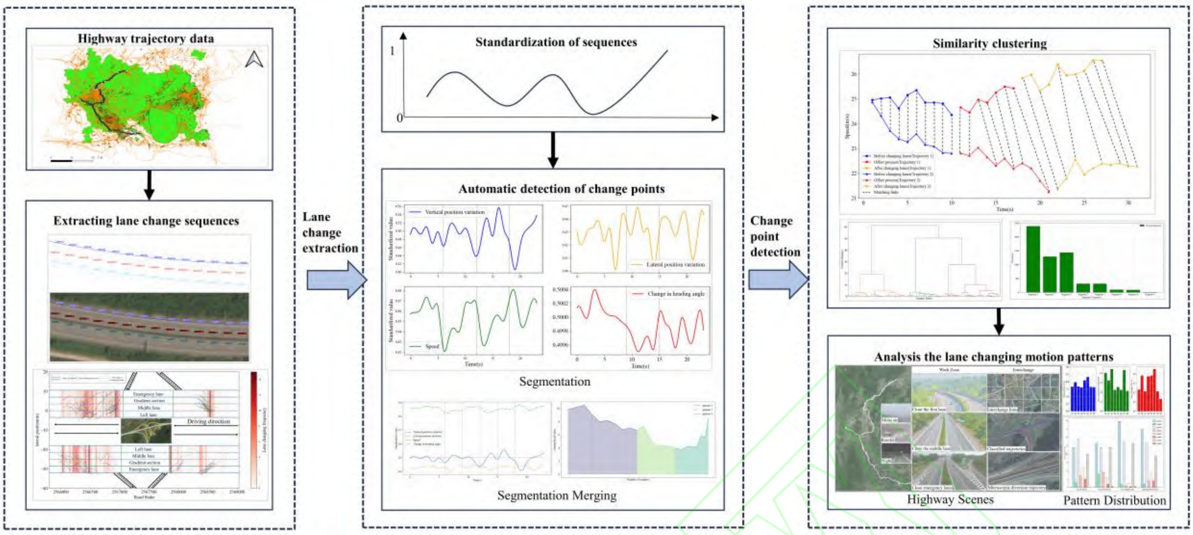


Fig 1. A Method Framework for Identifying Lane Changing Trajectory Patterns.

2.1. Change Point Detection Method

The change point detection method includes three parts: cost function, search method, and change point quantity constraint (Guijo-Rubio et al. 2021). The (PELT) algorithm can be considered an improvement on the optimal segmentation method (Haynes et al. 2017). Fundamentally, it operates using the same principle of minimizing the cost function:

$$C(\tau) = \sum_{i=1}^{m+1}[C(y_{\tau_{i-1+1}}) + \beta] \tag{1}$$

among τ It is a change point, and m is the total number of change points.

One of the challenges related to the change point detection of LC sequence is the need to monitor several key performance indicators and to automatically determine their parameters through training. The processing of multidimensional data needs to be standardized, which is called $LC_{1:n}^{i} = (lc_1^i, \ldots, lc_n^i)$ to represent the LC sequence identified in the ith dimension, where n is the sample size. The outliers of LC sequence are corrected by truncation method (Yarlagadda and Pawar 2022) and normalized by Equation (2).

$$\bar{LC}_k^i = \frac{LC_k^i - \bar{LC}^i}{std_{LC^i}}, \quad i = 1, \ldots, n \tag{2}$$

In the formula, $\overline{LC^{i}}$ and $std_{LC^{i}}$ are the average and standard deviation of the ith dimensional LC sequence, respectively; $\overline{LC_{k}^{i}}$ is the normalized value of a certain feature of LC at time k.

The setting of penalty conditions is the key to determining the accuracy of LC segment recognition, and it is necessary to avoid the local optimal impact of penalty values on recognition accuracy. Therefore, different data should be used for automatic penalty settings for LC scenarios in different environments. Here, we propose the following solution, referred to as APELT in the following text. It is worth noting that this process uses a massive number of variable point and LC sequences marked by experts in various dimensions to train the globally optimal beta, which is a continuous iterative process. For each LC, use the ED-PELT algorithm to generate boundary values for penalty values for multiple detectable change points in the LC sequence. Then we extract the upper and lower boundaries of the penalty values for the change points identified by the ED-PELT algorithm within the range of 20 points marked by experts. Finally, we decompose the maximum to minimum range of the beta values we extract into multiple values with a step size of 1, and use each decomposed value to traverse each boundary value to obtain a score, and select the beta value with the highest score as the global optimal value. Table 1 shows the steps of the algorithm:

The setting of penalty condition is the key to determine the recognition accuracy of LC points, and it is necessary to avoid the local optimal influence of penalty value on the recognition accuracy. Therefore, for LC behaviors under different road geometries, data from different scenes should be used for training to automatically obtain the penalty value. Here, we propose the following solutions, hereinafter referred to as APELT. It is worth noting that this process uses a large number of change points and LC sequences marked by experts in various dimensions to train the global optimal beta value, which is a continuous iterative process. For each LC, the ED-PELT algorithm is used to generate boundary values for the penalty values of multiple detectable LC points in the LC sequence. Then, we extract the upper and lower bounds of the penalty value of the change point within the range of 30 points before and after the expert mark point. Finally, we decompose the maximum to minimum range of the extracted beta value into multiple values with a step size of 1, and use each decomposed value to traverse each boundary value to obtain a score. Finally, we choose the beta value with the highest score as the global optimal value. Table 1 shows the pseudo code of the algorithm:

Table 1. The algorithm steps of APELT.

Algorithm 1 APELT procedure steps

```
Sequence ←lane changing sequence data $LC = [LC^1, LC^2, LC^3, LC^4]$
  annotated ←annotated sequences $AN_{LC} = [AN_{LC^1}, AN_{LC^2}, AN_{LC^3}, AN_{LC^4}]$
  $cp^1, \beta^1, cp^2, \beta^2, cp^3, \beta^3, cp^4, \beta^4$←GET_CP(ED-PELT($LC^1$),
  ED-PELT($LC^2$),ED-PELT($LC^3$), ED-PELT($LC^4$))
  $\beta$←$[\beta^1, \beta^2, \beta^3, \beta^4]$; $CP$ ←$[cp^1, cp^2, cp^3, cp^4]$
  for $i$ ← 0 to len(Sequence) - 1 do
    for $j$ ← 0 to len($CP[i]$) - 1 do
      if ABS(MIN(annotate[$i$], KEY=lambda x: ABS($x$ - $CP[i][j]$)) - $CP[i][j]$) < 10 then
        $\beta$_list.append($\beta[i][j]$)
  $\beta$_range ← [range(min($\beta$_list[0]), max($\beta$_list[0]) - 1),range(min($\beta$_list[1]), max($\beta$_list[1]) - 1)
    ,range(min($\beta$_list[2]), max($\beta$_list[2]) - 1),range(min($\beta$_list[3]), max($\beta$_list[3]) - 1)]
  for $i$ ← 0 to len($\beta$_range) - 1 do
    for $j$ ← 0 to len($\beta$_range[$i$]) - 1 do
      $\beta$_value ← 0
      for $k$ ← 0 to len($\beta[i]$) - 1 do
        if $\beta$_range[$i$][$j$] > $\beta[i][k]$[0] and $\beta$_range[$i$][$j + 1$] > $\beta[i][k]$[1] then
          $\beta$_value ←$\beta$_value + 1
      $\beta$_value_list.append($\beta$_value)
      Best_ $\beta$.append(argmax($\beta$_value_list))
      $\beta$_value_list ← []
  return Best_ $\beta$
```

APELT automatically searches for the comprehensive optimal penalty term according to the data from different environments, and its importance to trajectory data segmentation lies in simplifying the parameterization of change point detection. Engineers can use this scheme to annotate relatively short time series sets and then use them to calculate the optimal penalty. Another method is to establish the penalty value through trial and error, or to use the ED-PELT algorithm whose performance is lower than the program specification proposed in this paper. APELT method trains the penalty through a large number of LC data and LC points marked by experts. Therefore, it has the advantage that the trained model is more interpretable when identifying LCs, and there is no need to set a separate penalty for each LC. After that, LC with similar distribution can share the same penalty value. However, if the same penalty is used in different environments, such as work areas, different lanes and interchanges, it is best to retrain the corresponding penalty.

2.2. Similarity Clustering Method

DTW distance can be used to measure time series of unequal lengths, but its computational complexity is high and cannot meet the trigonometric inequality of distance measurement, often leading to excessive distortion. To address this issue, an improved distance measurement method based on similarity matching(SM) is proposed. It includes two main steps: distance measurement and distance matrix clustering.

2.2.1 Improved lane change point matching method

The improved method allows time series to have a certain degree of offset on the timeline. On the other hand, it satisfies the triangle inequality for distance measurement (Chen and Gao 2020). Given LC sequence $LC_i = \{lc_1, lc_2, \ldots, lc_m\}^i$, $lc_i = ([v_1^i, v_2^i, v_3^i, v_4^i], t_i, stage_i)$; if matching two sequences of unequal length LC_i and LC_j. It needs to meet the following two conditions to successfully match.

Condition 1 (Eq. (3)) is to satisfy the requirements of the trigonometric inequality by searching for all values within the range of t_j and match the most similar changes value t_i in each dimension:

$$t_i = \underset{t_i \in [t_j - \xi, t_j + \xi]}{argmin} \{\tfrac{1}{4}\textstyle\sum_{k=1}^{4}(\Delta v_k^i - \Delta v_k^j)^2\} \quad (3)$$

Condition 2 (Eq. (4)) is to limit the matching of trajectory points in different LC stages:

$$stage_i = stage_j \quad (4)$$

Among them, t_i represents the timestamp of LC_i: $[v_1^i, v_2^i, v_3^i, v_4^i]$ represents the four-dimensional eigenvalues, $stage_i$ indicates which stage of LC the trajectory at this moment belongs to. ξ is a threshold used to control the time allowed for axis movement. For point i of LC stage 1 (the nearest point is j); If the matching point of $i-1$ point is $j-k$ point, then the difference between the number of remaining points after $j-k$ point and the data amount of remaining points after i point is taken as the threshold ξ. That is to say, when a point has optional conditions, it preferentially selects the point with the smallest distance in the same stage as the matching point.

In Fig 2, the points marked with circles and triangles are similar matching points. Compared with point to point matching, this matching method improves efficiency and does not match points from different stages together; In addition, if there are multiple points that satisfy the condition when searching for a matching point, the point with the most similar fluctuation degree will be matched, and the peak will match the peak, unless there are no other values within the range that meet the condition.

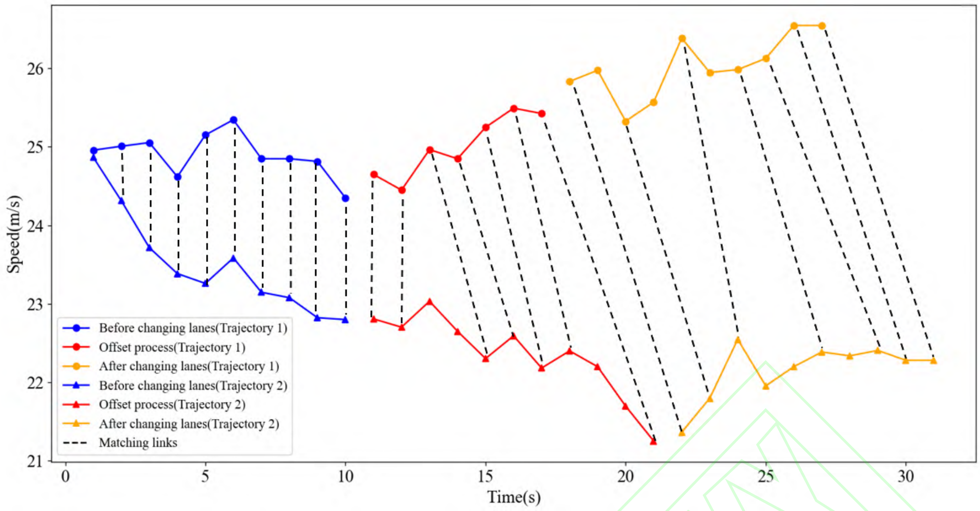


Fig 2. Schematic diagram of improving similarity matching.

2.2.2 Multidimensional distance calculation

After similarity matching, we need to calculate the four-dimensional distance between two fluctuation point sequences. Firstly, calculate the distance between each pair of matching points using Eq. (5).

$$d(lc_i, lc_j) = \sqrt{\frac{1}{4}((v_1^i - v_1^j)^2 + (v_2^i - v_2^j)^2 + (v_3^i - v_3^j)^2 + (v_4^i - v_4^j)^2)} \tag{5}$$

Then, calculate the three consecutive points lc_{i-1}, lc_i, lc_{i+1}.The fluctuation level of (i+1):

$$VD_{lc_i} = \frac{1}{4}\sum_{k=1}^{4}\frac{|v_k^{i+1}-v_k^i|+|v_k^i-v_k^{i-1}|}{2} \tag{6}$$

And $i = 2,3,\ldots,m, VD_{lc_1} = \frac{1}{4}\sum_{k=1}^{4}|v_k^2 - v_k^1|, VD_{lc_m} = \frac{1}{4}\sum_{k=1}^{4}|v_m^2 - v_{m-1}^1|$

Calculate the information weight IW and similarity matching degree SMD again (Chen and Gao 2020):

$$IW_{lc_i} = \frac{VD_{lc_i}}{\sum VD_{lc_i}} \tag{7}$$

$$SMD_{LC_i,LC_j} = \frac{1}{\sum|IW_{lc_i} - IW_{lc_j}|} \tag{8}$$

Finally, the distance between the two sequences is:

$$dist(LC_i, LC_j) = \frac{1}{SMD_{LC_i,LC_j}} \sum d(lc_i, lc_j) \tag{9}$$

Based on the concept of existing similarity matching techniques (Chen and Gao 2020), we have summarized an improved distance calculation method based on similarity matching (SM). The algorithm steps are shown in Table 2:

Table 2. The algorithm steps of SM.

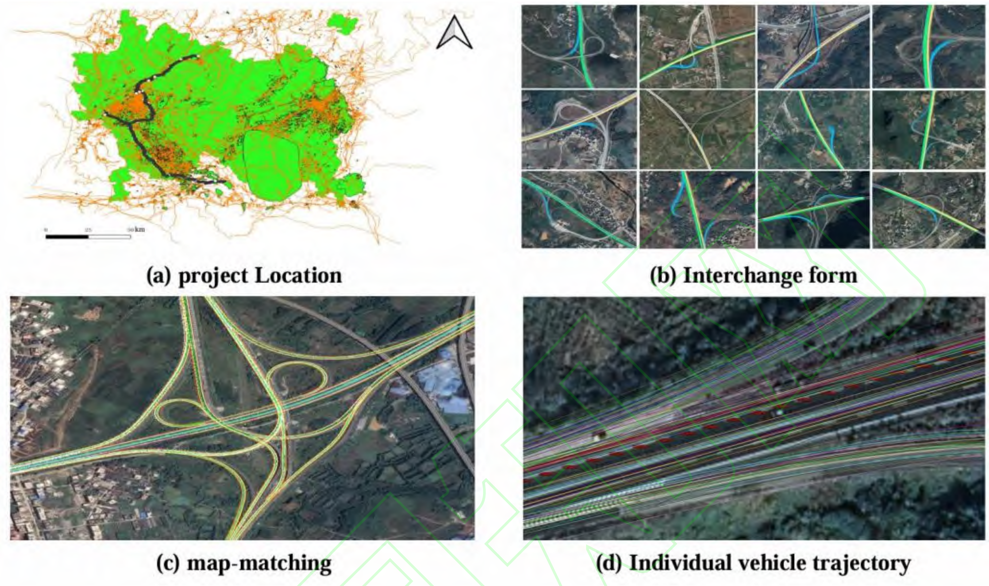
Algorithm 2 SM procedure steps

```
    $LC_i$, $LC_j$, $stage_{LC_i}$, $stage_{LC_j}$←Lane change sequence and stages
$VD_{lc_i}$ ← [$Eq5(LC_i[i-1], LC_i[i], LC_i[i+1])$ if $i$ > 0
and $i$ < len($LC_i$) - 1 else 0 for $i$ in range(len($LC_i$))]
$VD_{lc_j}$ ← [$Eq5(LC_j[j-1], LC_j[j], LC_j[j+1])$ if $j$ > 0
and $j$ < len($LC_j$) - 1 else 0 for $j$ in range(len($LC_j$))]
$t_i$ ← [argmin([$Eq5(LC_i[i], LC_j[j])$ if $stage_{LC_i}[i]$ == $stage_{LC_j}[j]$
else 9999 for $j$ in range(len($LC_j$))])
          for $i$ in range(len($LC_i$))]
    $\xi$ ← 0
    for $i$ ← 0 to len($t_i$) - 1 do
        if $t_i[i]$ - $\xi$ ≤$t_i[i]$ then
            $t_i[i]$ ←$t_i[i]$ - $\xi$
        else if $t_i[i]$ < $t_i[i]$ + $\xi$ then
            $t_i[i]$ ←$t_i[i]$+ $\xi$
    $IW_{lc_i}$ ← [$VD_{lc_i}[i]$ / sum($VD_{lc_i}$) for $i$ in range(len($VD_{lc_i}$))]
    $IW_{lc_j}$ ← [$VD_{lc_j}[i]$ / sum($VD_{lc_j}$) for $i$ in range(len($VD_{lc_j}$))]
$dist_{ij}$ ← [abs($IW_{lc_i}[i]$ - $IW_{lc_j}[t_i[i]]$) * $Eq5(LC_i[i],LC_j[t_i[i]])$
for $i$ in range(len($t_i$))]
    return $dist_{ij}$
```

3. Data Preparation

3.1. Descriptives

The data in this paper is the floating vehicle trajectory data of heavy vehicles, the frequency is 1HZ, the positioning accuracy is less than 1m, and the speed measurement accuracy is 0.2m/s, which is collected by more than 400 drivers driving heavy trucks on an expressway through on-board GPS, as shown by the black trajectory in Fig 3(a); Our data contains a variety of highway scenes, as shown in Fig 3(b) is a complex form of interchange. In addition, we use Python's "Pyautocad" library to display all the trajectory points on the CAD map, and then import them into the Ovi map, as shown in Fig 3(c). Our trajectory in a certain length is superimposed by the driving trajectories in different time periods, as shown in Fig 3(d). In short, our data covers a wide range of scenes in space, including days, nights, rainfall and other multi-time domains in time. The summary information of our data is shown in Table 3.


(a) project Location

(b) Interchange form

(c) map-matching

(d) Individual vehicle trajectory

Fig 3. Example of Full Time Domain Trajectory Big Data.

Table 3. Basic information of trajectory big data.

list	value	list	value
Road length	156km	Number of interchanges	18
Speed limit value	Inner 120 km/h Outer 100km/h	The number of work zone events	334
lane-width	3.75m	Time Span	3 months

3.2. LC events extraction

We provide a LC extraction method for full-time domain trajectory data. The road under study consists of three lanes, including a middle lane, an overtaking lane (innermost lane) and an emergency lane (outermost lane). We use kd-tree (Zheng et al. 2022) algorithm to match each trajectory point to the road pile number according to the road pile number information, as shown in Fig 4. Then we calculate the distance between each track point and the road centerline, so that we can determine the lane where each point is located according to the actual geometric information of the road and convert these points into Frenet coordinate system with the road centerline as the reference line. In addition, in order to minimize LC recognition errors, we have established several rules for LC detection, as follows:

- Step 1: Calculate the lateral offset of the continuous trajectory from the road centerline per second. Defines that the outward offset is negative and the inward offset is positive.
- Step 2: Sum the lateral continuous deviations, and extract these trajectory data when the cumulative lateral deviation exceeds the width of the vehicle.
- Step 3: Determine the lane number of the starting point and the lane number of the ending point for the driving process when the accumulated lateral deviation reaches the vehicle width in step 2. If the two lane numbers are different, they are classified as LC trajectory data.
- Step 4: For the collected LC data, the information related to time and space is matched and divided into mandatory, free and other LC types.

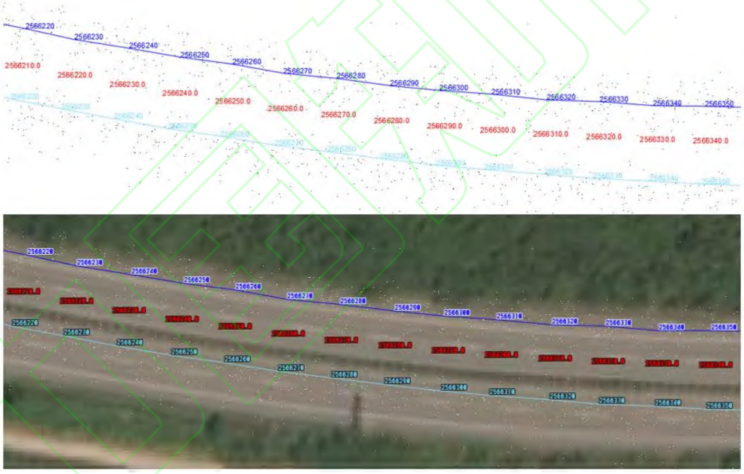


Fig 4. Road Stake System and Map Matching.

As shown in Fig 5, we show an example of extracting the mandatory LC behavior before the junction of an interchange. We divide the two-way road into four areas, including upstream and downstream diversion and confluence areas, and generate lane lines and ramp positions according to coordinate information. Each area shows the frequency of LC starting position by thermal map, and the trajectory lines of different colors show LC offset sequence. The figure shows that the frequency and position of LC are different in different geometric scenes. In addition, the road scenes included in our study are as shown in Fig 6, including three LC scenarios under workspace events in addition to interchanges; Finally, we divide the length extraction of the LC process into three stages: 10 seconds before the start, the LC offset process and 10 seconds after the LC.

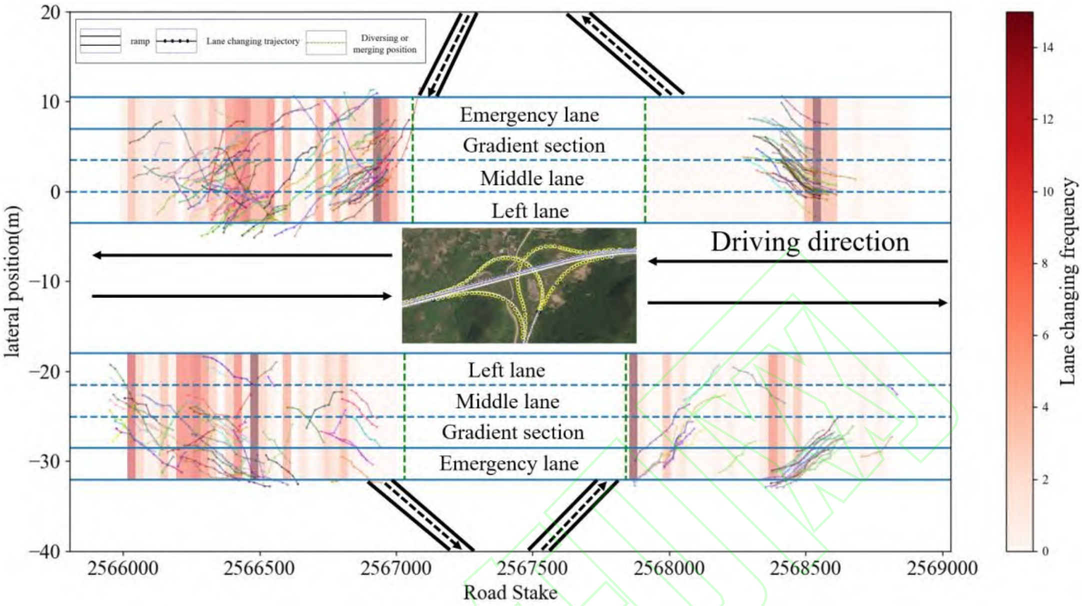


Fig 5. Lane changing extraction and frequency distribution in interchange areas.

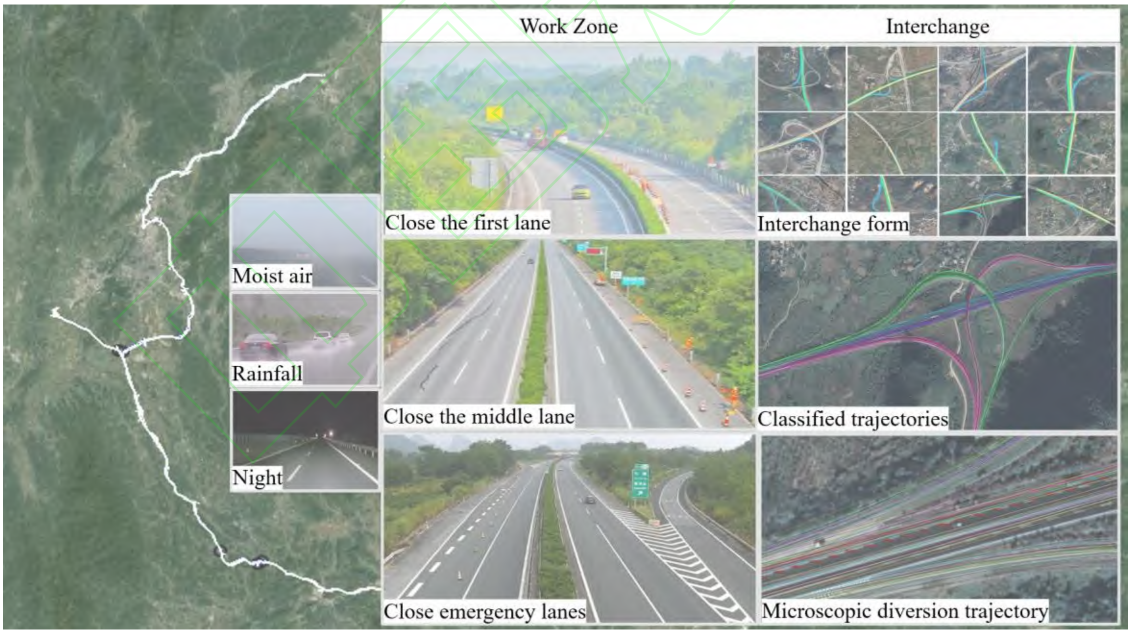


Fig 6. The geometric types of highways covered by data.

3.3. evaluation criterion

Once the algorithm performs clustering, it is necessary to verify or evaluate the clustering results. We evaluate the change point detection algorithm using F1 score and accuracy. For similarity clustering method, the number of hierarchical clusters is first determined based on the contour coefficient, and then the Kappa coefficient can be used to check whether the most similar fragments of a certain fragment are classified into the same category.

3.3.1 Silhouette Function

The Silhouette function (SF) (Hamedi and Shad 2022) is considered an efficient method for analyzing the separation distance between clusters, especially in the absence of dataset class labels. The SF graph provides a visual measurement within the range of $[-1,1]$, which evaluates the closeness of patterns within galaxy clusters, with observations with poor clustering often approaching -1 and observations with good clustering approaching -1. When the contour value is greater than 0.5, it indicates good clustering results. When the silhouette value approaches 0, it indicates that the sample is located near or above the determining boundary between two adjacent galaxy clusters. Negative values indicate that these samples may have been mistakenly clustered together. The contour value can be defined as:

$$S(i) = (b_i - a_i)/ max(b_i, a_i) \tag{10}$$

In the equation, b_i represents the average distance from the observed value i to the nearest neighboring observations within the same galaxy cluster, a_i represents the mean of all other observations within the same galaxy cluster from observation result i.

3.3.2 Precision of Similarity Calculation

In the context of machine learning, F1 score and accuracy are two common indicators for evaluating the performance of classification models, which can evaluate the performance of sequence segmentation methods. In addition, as a statistical coefficient, Cohen's kappa represents the level of reliability and accuracy of statistical classification (Hamedi and Shad 2022). Similar to the correlation coefficient, Cohen's kappa can be between 1 and+1, where 0 represents the consistency of random opportunity expectations and 1 represents complete consistency among raters. Like all other relevant statistical data, Cohen's kappa is a standardized value, therefore it has been explained in the same way in various surveys. Cohen proposed the following explanation for Kappa values: values between 0.81 and 1.00 represent almost identical values. Firstly, we must generate a confusion matrix, which is a unique table layout that allows for visualizing the behavior of algorithms, typically supervised learning algorithms (commonly referred to as matching matrices in unsupervised learning). The entire row of the matrix represents instances in the actual class, while the entire column represents examples in the expected class, and vice versa. Cohen's kappa can be calculated using the following equation: K represents the clustering of LC process indicators to obtain the labels for each LC process. Subsequently, traverse the entire LC database to search for the most similar trajectory Si for each LC process in the similarity matrix. Then check if the category of Si is consistent with it. The calculation equation is as follows:

$$K = (P_0 - P_e)/(1 - P_e) \tag{11}$$

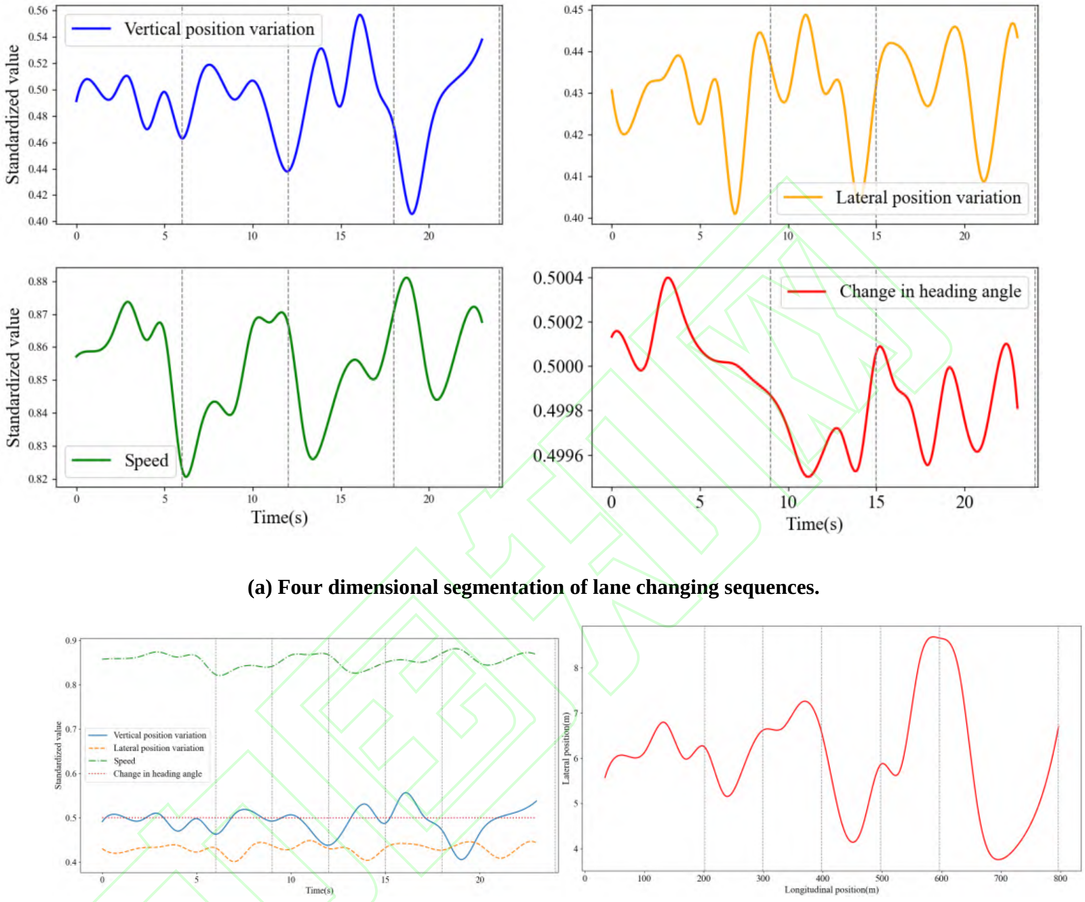
in which P_e stands for the hypothetical probability of chance agreement, and P_0 represents the relative observed agreement amongst raters. Using Cohen's kappa score, the performance of Similarity measurement has been estimated in the results section.

4. Experiment Results

4.1. LC sequence segmentation

In this section, the experimental results of LC sequence are given by APELT algorithm, and compared with the algorithms (GMM-HMM (Zhang et al. 2023), HDP-HMM (Chen et al. 2022) and sliding window (Qi et al. 2022)) in LC sequence segmentation. In addition, APELT algorithm has been used to delete the data of abnormal trajectory values and not to delete the data of abnormal trajectory values (tracks that deviate from the road edge or are interrupted by signals), and automatically select the penalty item to compare with ED-PELT. The algorithm trains the global optimal parameter beta, so that the minimum segment length (the minimum distance between two change points) is equal to 30 points, which is equivalent to 3 seconds (the accuracy is 0.1s through linear interpolation). The purpose of the experiment is to evaluate the efficiency and accuracy of the algorithm in LC sequence application. As far as we know, this is the first time to apply change point detection to the report of real LC trajectory.

First of all, we choose velocity v, lateral acceleration d, longitudinal acceleration s and heading angle b as input variables. The first three variables represent the stability of the LC trajectory, and the change of heading angle is related to the driving of the vehicle on the curve section and the operation of the steering wheel. First, divide the standardized variable sequence of each dimension in sequence, as shown in Fig 7(a), and then take the union of the change points of each divided dimension, as shown in Fig 7(b). Although multiple segments are segmented, subsequent similarity clustering algorithms can merge similar segments in this step (Fig 7(c)).

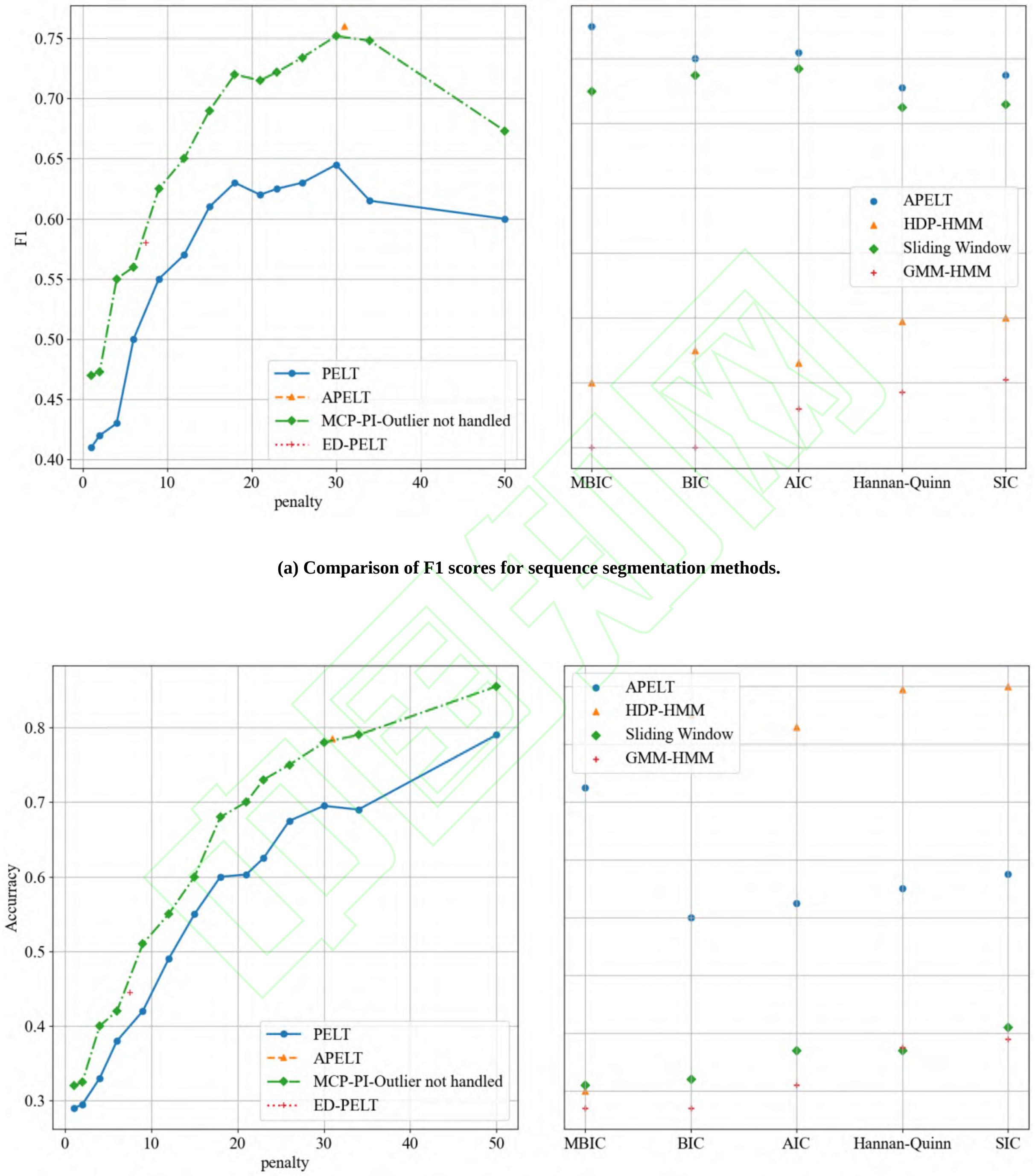


(a) Four dimensional segmentation of lane changing sequences.

(b) Segmentation of lane changing sequences **(c) Merge lane changing trajectories of fragments.**

Fig 7. Schematic diagram of lane changing sequence segmentation. (a). (b). (c).

The accuracy scores of different algorithms are shown in Fig 8. The results obtained from the penalty values arbitrarily selected in the PELT algorithm (with and without abnormal trajectory removal) and the PELT results when using two optimization schemes (namely ED-PELT and APELT) for penalty search are displayed in the left panel. One point represents two schemes, and each algorithm calculates its own penalty value. The right panel compares the results produced by GMM-HMM, HDP-HMM, sliding window and APELT with different information standards.



(a) Comparison of F1 scores for sequence segmentation methods.

(b) The accuracy of comparing sequence segmentation methods.

Fig 8. Evaluation of MCP-PI method. (a). (b).

In addition, as shown in Fig 9, the statistical information of segmentation points generated when using different segmentation algorithms on the test set shows that the number of segments when using HDP-HMM and GMM-

HMM is much higher than that when using APELT, and their longest segment duration is much lower, which is obviously unrealistic. In addition, when using HDP-HMM and GMM-HMM methods, the positions of segmentation points before and after LC are concentrated, and the results of GMM-HMM segmentation are more dispersed. Due to the automatic training of global optimal parameter values, when APELT is used, the segmentation points are more focused on the beginning and end of LC, and the duration of the segmentation segment will not be abnormally low. The above segmentation positions indicate that HDP-HMM and GMM-HMM are more sensitive to the changing trend of multidimensional sequences, which may not be suitable for short sequence segmentation. In addition, because the state durations of HDP-HMM and GMM-HMM are modeled by Poisson distribution, many segments are divided into time series with a length of 0.1 seconds, which is meaningless for pattern extraction, because the duration of LC behavior patterns should not be too long or too short (Chen et al. 2023). In addition, as shown in Fig 9, the statistical information of segmentation points generated when using different segmentation algorithms on the test set shows that the number of segments when using HDP-HMM and GMM-HMM is much higher than that when using APELT, and their longest segment duration is much lower, which is obviously unrealistic. In addition, when using HDP-HMM and GMM-HMM methods, the positions of segmentation points before and after LC are concentrated, and the results of GMM-HMM segmentation are more dispersed. Due to the automatic training of global optimal parameter values, when APELT is used, the segmentation points are more focused on the beginning and end of LC, and the duration of the segmentation segment will not be abnormally low. The above segmentation positions indicate that HDP-HMM and GMM-HMM are more sensitive to the changing trend of multidimensional sequences, which may not be suitable for short sequence segmentation. In addition, because the state durations of HDP-HMM and GMM-HMM are modeled by Poisson distribution, many segments are divided into time series with a length of 0.1 seconds, which is meaningless for pattern extraction, because the duration of LC behavior patterns should not be too long or too short (Chen et al. 2023).

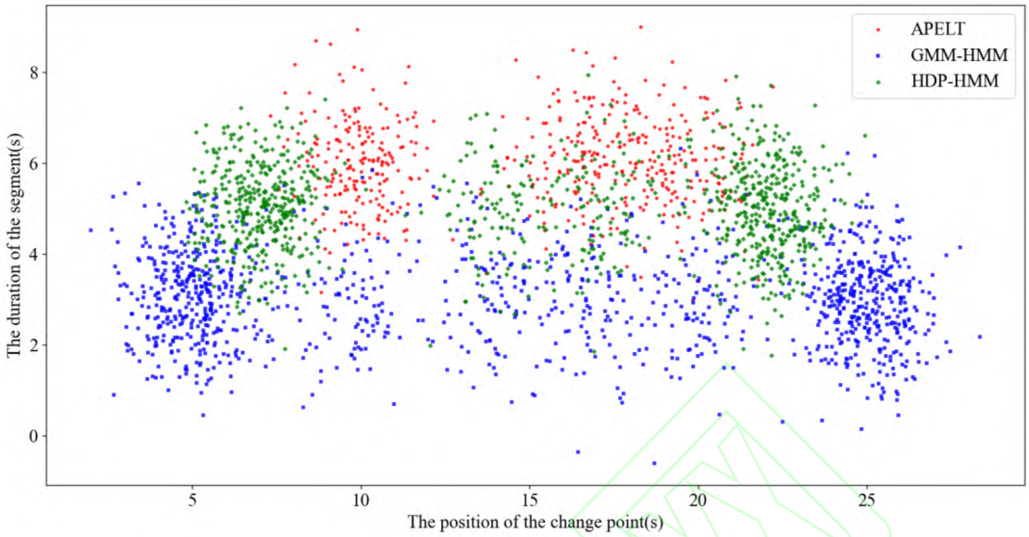


Fig 9. The position of change points and the length distribution of segments.

In APELT algorithm, firstly, it is positive to remove the influence of outliers, that is, when the structure of time series pattern is destroyed, this behavior can optimize the selection of penalty values and minimize the error of finding time points. As can be seen from the figure, it improves F1 and accuracy. For example, when the penalty value is 10, the improved algorithm improves the F1 score by about 8%; When the penalty value is about 30, the improved algorithm shows the highest F1 score. The accuracy results are similar, and APELT is obviously superior to ED-PELT in F1 and accuracy. In addition, compared with most manually selected penalty values, the results obtained by using the information standard in penalty items are worse, because the F1 sum of manually set penalty items is not accurate, and the segmentation position is not as realistic as the penalty items generated by automatic training.

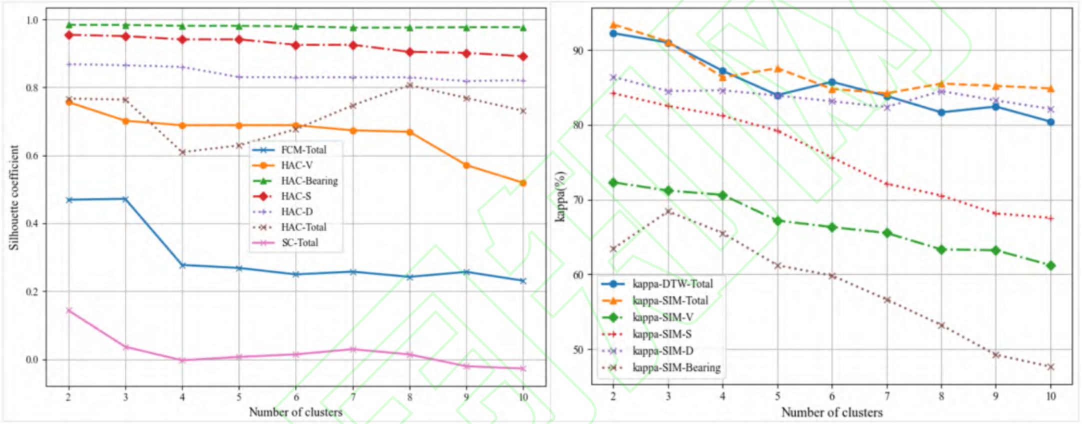
4.2. Similarity clustering of LC fragments

We calculate the similarity matrix of each dimension and the similarity matrix of four-dimensional comprehensive distance, and use spectral clustering, fuzzy C-means clustering (FCM) and hierarchical clustering (HAC) to evaluate the effectiveness of our fragment clustering. In the clustering step, it is necessary to set the number of clusters for the similarity clustering model. For models using different segmentation and clustering algorithms, we set the number k of patterns in the range of 2 to 10. In the clustering algorithm using similarity matrix shown in Fig 10(a), the variation value of contour coefficient of HAC method is significantly higher than that of FCM and SC method, which shows that HAC has the best clustering effect on LC similarity matrix. However, the contour coefficient of HAC method for clustering one-dimensional variables is greater than that of multi-dimensional clustering, because one-dimensional data is clearer and more compact, and the overall matrix is four-dimensional, which may introduce noise or fuzzy clustering boundaries. Eight categories with the highest contour coefficients (defined as S1~S8) are selected as the number of clusters.

Next, we will test the results of similarity measurement based on the difference in duration between the fragment and its most similar fragment. Firstly, we calculate a confusion matrix (Hamedi, Shad, and Jamali 2023), which is a table used to determine the performance of classification algorithms. The confusion matrix for LC trajectory clustering is shown in Table 4. Among the 5458 trajectories classified in cluster number 1 before LC, 4830 trajectories have been correctly predicted. This means that 4830 trajectories and their most similar trajectories were correctly placed in the same category, while 638 most similar trajectories were mistakenly placed in other clusters. For the entire cluster, the matrix is subject to similar control. Then, based on Eq. (11), calculate the Kappa coefficient to characterize the accuracy of similarity measurement, as shown in Fig 10(b). There are a total of 9604 LC sections, among which our SIM-Match clustering has a Kappa score greater than 0.8, which is better than DTW. This is because our algorithm restricts the matching conditions to make the duration of each segment closer to the duration of the most similar segment. In addition, we consider the fluctuation characteristics of the data to match points, which increases the distance between LC segment groups and makes the clustering effect more obvious. When the number of clusters exceeds the number we choose, the accuracy decreases. In addition, the kappa score of four-dimensional weighted matrix clustering is higher than that of single dimensional matrix clustering, which reflects that the duration of a segment is not determined by a single dimensional pattern such as longitudinal velocity, but is more closely related to the complex interaction of multiple dimensions of patterns such as lateral acceleration and heading angle during LC. This indicates that our similarity measurement accuracy is very high, and finally, we need to check the effectiveness of the matrix clustering method. We traversed each LC and checked whether each LC trajectory fragment and its most similar fragment (expected by the similarity clustering algorithm) were included in the same class. The result showed that the accuracy of the HAC method was as high as 92.6%, higher than that of SC and FCM techniques, indicating that HAC is more suitable for clustering multidimensional matrices.

Table 4. Confusion Matrix of HAC Clustering Results.

		Predicted class								
	class	1	2	3	4	5	6	7	8	sum
Observed class	1	4830	148	161	29	163	27	97	3	5458
	2	80	179	25	9	21	5	29	0	348
	3	138	57	1391	15	56	17	74	1	1749
	4	27	3	5	43	4	1	5	0	88
	5	115	21	14	2	219	4	11	0	386
	6	3	7	4	2	3	40	3	1	63
	7	109	53	83	11	46	9	1191	0	1502
	8	6	0	1	1	0	1	1	0	10
	sum	5308	468	1684	112	512	104	1411	5	9604



(a) (b)

Fig 10. The variation of evaluation coefficient with the number of clusters.(a) Silhouette coefficient. (b) Kappa score.

5. Discussion

5.1. Descriptiveness of the Extracted Behavioral Patterns

In this paper, the LC trajectory pattern we extracted not only needs to have good performance in quantitative indicators, but also needs to have interpretable characteristics. Therefore, we analyzed the trajectory characteristics of eight LC patterns, and then investigated the frequency distribution of free and mandatory LC types in different road scenes. Among them, free LC mainly means that vehicles are driving in straight sections with good geometric conditions and are not affected by events, while mandatory LC includes LC behavior of vehicles at the exit ramp of interchange due to diversion and LC behavior of roads due to construction of closed lanes.

For this reason, we first count the frequencies of all LC trajectory patterns, as shown in Fig 11, in which pattern 1 is the most common, accounting for 40.36% of the whole, and the average of its lateral and longitudinal acceleration in this pattern is -0.01 m/s^2, and the duration is 9.2s, which is the longest. Because the LC sequence includes the sequence before and after LC besides the LC offset process, compared with other patterns, it belongs to the normal driving state of LC process (it does not involve the vehicle's lane shift and acceleration and deceleration in the longitudinal direction); The frequencies of pattern 7 and pattern 3 are 21.79% and 24.81%, respectively. Although they have no obvious change in lateral acceleration (-0.01 m/s^2, 0.03 m/s^2), they respectively show acceleration pattern (0.68 m/s^2) and deceleration pattern (-0.74 m/s^2) in longitudinal speed, and their duration is 5.31 s . In addition, pattern 2 and pattern 5 account for 5.23% and 5.21% respectively, which are contrary to the longitudinal acceleration and deceleration patterns, and their average lateral accelerations are 2.01 m/s^2 and -2.07 m/s^2 respectively, which are stable in the longitudinal speed, belonging to LC offset accelerations in two different patterns, and their duration is shorter than that in other patterns (4.63s and 4.82s). pattern 4 (1.42%), pattern 6 (1.42%) and pattern 8 (0.14%) are relatively few. The lateral movement of pattern 4 is more intense (lateral acceleration -5.31 m/s^2) and the duration is 1.49s shorter than that of pattern 5. Compared with pattern 2, pattern 6 is not only more intense in lateral acceleration (5.40 m/s^2), but also has a deceleration movement in longitudinal speed (-0.18 m/s^2), a significant change in heading angle (2.59 degrees) and the shortest duration of 3.2s However, the heading angle of pattern 8 changes the most, which may be related to the vehicle changing lanes on the curve and steering wheel manipulation greatly. Fig 12 reveals the interesting phenomenon of transition between different patterns. Specifically, all patterns are switched from pattern 1 (normal driving) with higher frequency. Except pattern 1, most patterns have the highest self-conversion frequency, which indicates that these trajectory patterns generally appear continuously in LC. Pattern 4 (transverse severe deceleration pattern) is more converted from transverse deceleration pattern, transverse acceleration pattern and longitudinal deceleration pattern, while pattern 6 (transverse severe acceleration pattern) is more converted from transverse deceleration pattern, transverse acceleration pattern and longitudinal acceleration pattern.

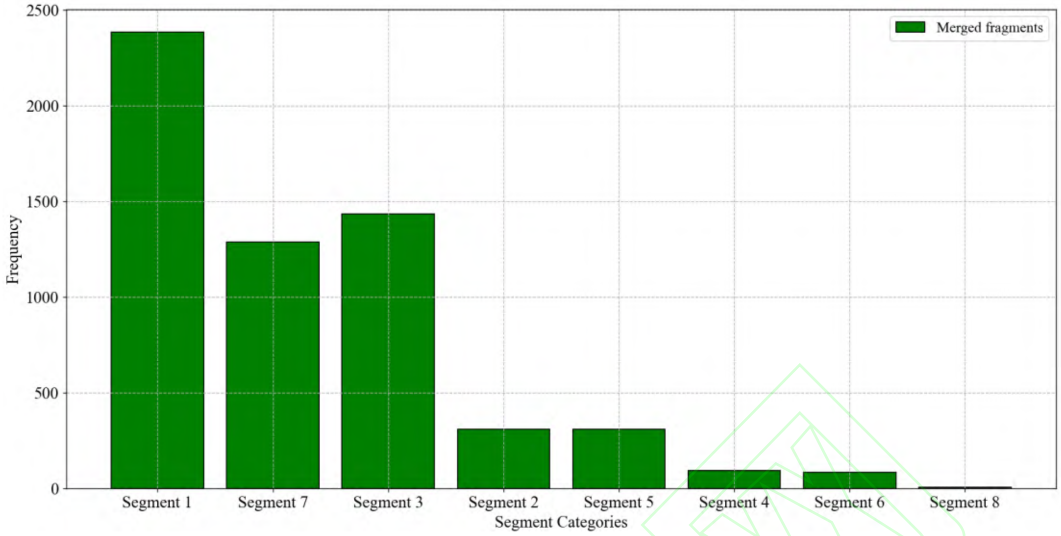


Fig 11. Frequency of lane changing patterns.

Then it is verified according to the statistical indicators (mean, median and standard deviation) of each feature, as shown in Table 6. The results show that the statistical values of different LC trajectory patterns are different, and their average values are significantly different. As shown in Table 5, the p values are all less than 0.05.

Table 5. The significance test results of analysis indicators.

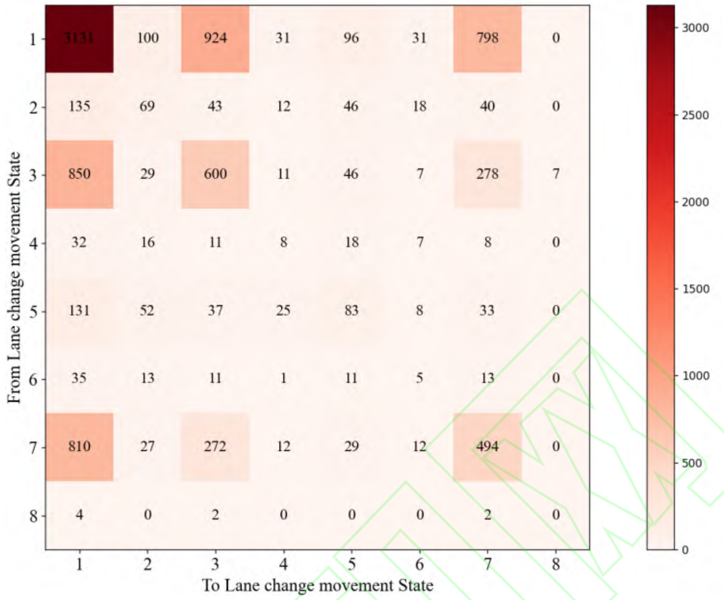
	$\boldsymbol{v}$	$\boldsymbol{d}$	$\boldsymbol{s}$	$\boldsymbol{b}$	$\boldsymbol{t}$	$\boldsymbol{wet}$	RF	$Night$
F-statistic	25.410	5872.806	1399.814	1559.584	247.405	3.007	3.356	17.051
p-value	0.000	0.000	0.000	0.000	0.000	0.004	0.001	0.000

*Statistically significant (p < 0.005).

Table 6. Statistical characteristics of analysis indicators.

Segment Type		$S1$	$S2$	$S3$	$S4$	$S5$	$S6$	$S7$	$S8$
$\boldsymbol{v}$	mean	24.23	25.24	24.52	21.53	22.92	27.15	24.33	25.32
	std	3.49	3.30	3.69	5.09	3.99	3.44	3.87	1.99
	median	24.07	25.18	24.27	22.12	23.48	27.20	24.19	24.66
$\boldsymbol{d}$	mean	-0.01	2.01	-0.01	-5.31	-2.07	5.40	0.03	0.47
	std	0.26	0.60	0.36	1.39	0.63	1.60	0.41	0.62
	median	0.00	2.26	0.00	-4.88	-2.29	4.90	0.01	0.34
$\boldsymbol{s}$	mean	-0.01	0.00	0.68	0.03	0.06	-0.18	-0.74	-0.94
	std	0.18	0.43	0.34	0.93	0.56	1.04	0.52	1.49
	median	0.00	0.00	0.59	0.00	0.05	-0.03	-0.60	-1.15
$\boldsymbol{b}$	mean	0.70	1.11	0.89	1.92	1.37	2.59	1.15	144.69
	std	2.63	5.23	3.19	9.72	4.70	9.78	3.68	44.87
	median	0.50	0.45	0.50	0.46	0.53	0.45	0.50	179.00
$\boldsymbol{t}$	mean	9.20	4.63	5.66	3.33	4.82	3.20	5.31	3.25

std	5.15	2.37	2.72	1.08	2.44	0.74	2.67	0.66
median	9.00	3.00	6.00	3.00	3.00	3.00	5.00	3.00



	1	2	3	4	5	6	7	8
1	3131	100	924	31	96	31	798	0
2	135	69	43	12	46	18	40	0
3	850	29	600	11	46	7	278	7
4	32	16	11	8	18	7	8	0
5	131	52	37	25	83	8	33	0
6	35	13	11	1	11	5	13	0
7	810	27	272	12	29	12	494	0
8	4	0	2	0	0	0	2	0

Fig 12. Pattern conversion frequency matrix.

5.2. Distribution of lane change patterns at different times

We investigated the distribution of different models in different air humidity, rainfall and at night, as shown in Fig 13. The humidity of our data refers to the air humidity of about 1.25~2 meters on the ground; Rainfall refers to the depth of liquid or solid (after melting) water falling from the sky to the ground on the horizontal plane, and night refers to the time period from 8: 00 pm to 5: 00 am the next day. The influence of air humidity on different patterns is not significant; However, the greater the rainfall, the more vertical acceleration and deceleration patterns appear; This shows that in rainy days, the friction on the road surface decreases and the driving of vehicles is more affected. In this case, the driver needs to control the vehicle more carefully to avoid sudden acceleration or deceleration. Compared with the daytime, there are more lateral acceleration and violent lateral acceleration patterns at night, which is because the visibility at night is poor, which further affects the stability of LC driving.

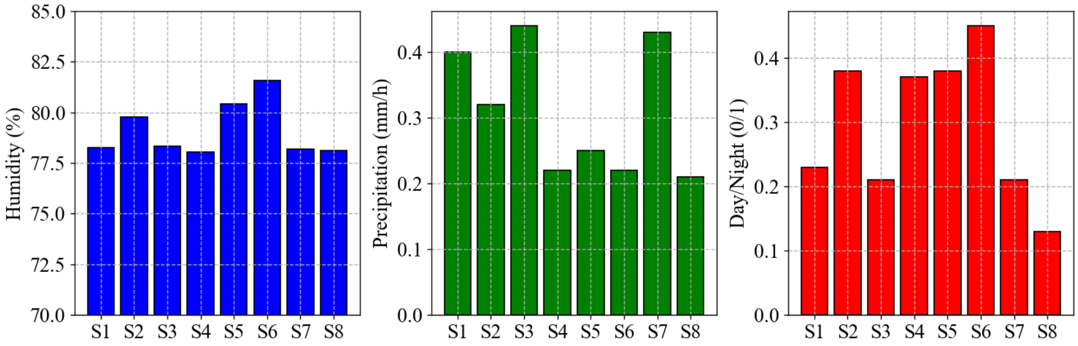
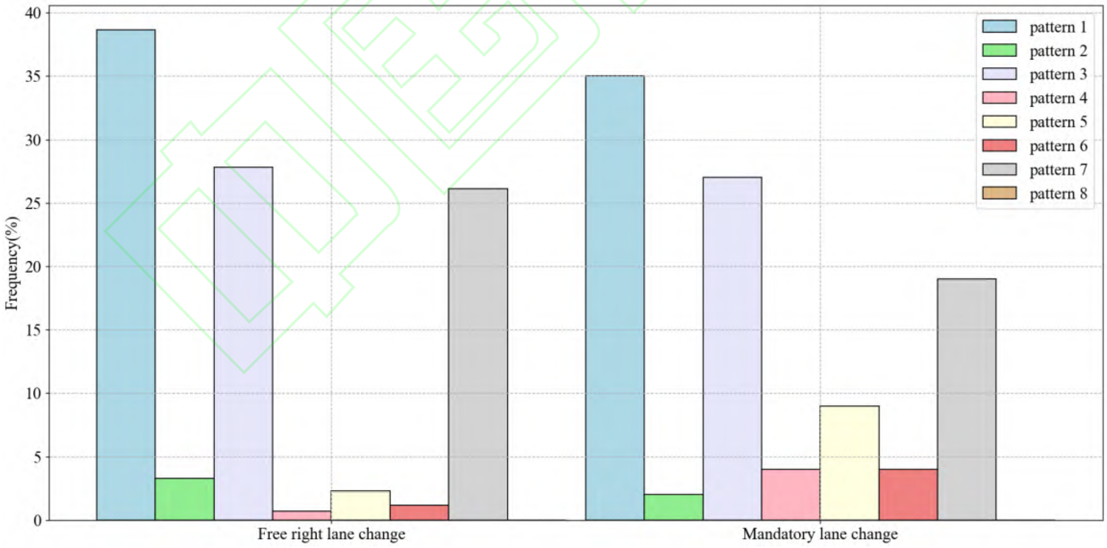
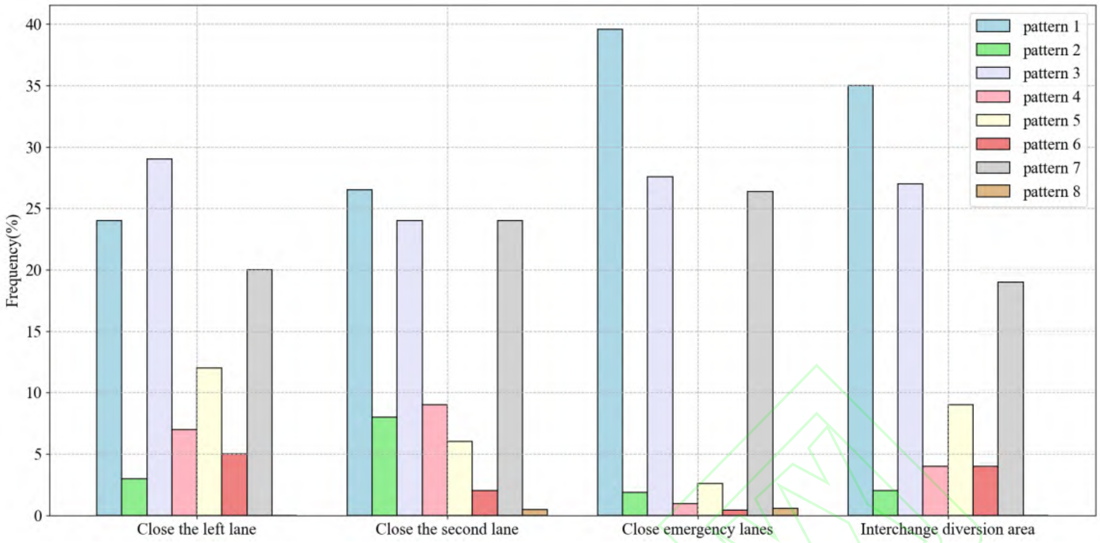


Fig 13. The influence of road surface humidity, rainfall, and night on lane changing patterns.

Because the LC trajectory may behave differently in different road geometric scenarios, we further investigated the pattern distribution of the scene as shown in Fig 14. Fig 14(a) compares the pattern distributions of free LC and mandatory LC. The mandatory LC is obviously more than free LC in pattern 4 (severe lateral deceleration), Pattern 5 (severe lateral deceleration) and pattern 6 (severe lateral acceleration), indicating that drivers are more inclined to take more hasty actions when mandatory LC. However, the acceleration offset pattern in pattern 2 is less than that in free LC, which indicates that drivers are more inclined to seek comfortable and stable driving conditions in the case of free LC.



(a) Comparison between free and mandatory lane changing.



(b) Comparison of lane changing between interchanges and work areas.

Fig 14. Frequency distribution of lane changing. (a). (b).

Finally, we compare the distribution of mandatory LC patterns between the working area with three lane types of closed roads and the exit area of interchange. Among them, the closed emergency lane has the least influence on lane changing, which is almost similar to the free LC of straight road sections. The closed left lane is similar to the mandatory LC of interchange, but the lateral rapid deceleration and lateral deceleration pattern are higher than that of interchange. Compared with the first lane, the closed middle lane shows more lateral acceleration patterns and severe lateral deceleration patterns; On the whole, the closure of the left lane, the working area of the middle lane and the diversion area of the interchange have a great influence on the LC pattern, and these findings have certain reference significance for the geometric design of the road.

6. CONCLUSION

In this paper, APELT-SM, a method framework for extracting LC trajectory patterns of full-time domain trajectory big data, is proposed, and LC sequences are segmented by improved change point detection technology. APELT can consider LCs in different highway geometric scenes to train the optimal penalty value; In addition, considering the excessive distortion of DTW algorithm, we improve the similarity distance measurement method and propose a similarity matching method SM to cluster sub-segments. We compare these two parts with the most advanced research methods; Finally, we discuss the characteristics and distribution of the recognized LC trajectory patterns to further evaluate the interpretability of our fragment types. In fact, the proposed method framework shows very accurate results in LC trajectory pattern recognition, which provides a novel idea for investigating the influence of expressway geometric design elements on LC trajectory. The results are summarized as follows:

APELT can help to consider the expressway scene and LC stage in LC pattern recognition, thus improving the accuracy. It is worth mentioning that the position and quantity of change point detection are more realistic, because this technology is a segmentation technology suitable for short sequences. The improved penalty optimization results also show that APELT provides a good result, which proves that the proposed method can successfully segment the LC trajectory sequence. Finally, the performance of APELT algorithm is related to the change points marked manually, which is very flexible. Compared with the existing segmentation techniques, our proposed change point detection technique has higher F1 score and accuracy in segmentation of LC sequences.

Our improved clustering method SM based on similarity matching can avoid point matching in different LC stages, such as LC preparation and execution, and set the position constraint of point-to-point matching in the matching process. This can effectively solve the problem of excessive distortion in distance measurement; According to the hierarchical clustering results, Kappa score is > 0.8, and the improved algorithm similarity clustering method has played an effective role in identifying LC trajectory patterns.

Finally, we further discuss the frequency and duration of the pattern and the conversion probability between different patterns, and can divide the LC trajectory pattern into the LC process and the lateral or longitudinal violent acceleration and deceleration trajectory pattern before and after LC deviation. An interesting discovery is that pattern 4 (violent lateral deceleration pattern) is more converted from lateral deceleration pattern, lateral acceleration pattern and longitudinal deceleration pattern, while pattern 6 (violent lateral acceleration pattern) is more converted from lateral deceleration pattern, lateral acceleration pattern and longitudinal acceleration pattern. In addition, in time, the greater the rainfall, the more vertical acceleration and deceleration patterns appear; This shows that in rainy days, the friction on the road surface decreases and the driving of vehicles is more affected. Compared with the daytime, there are more lateral acceleration and severe lateral acceleration patterns at night, which is because the visibility at night is poor, which in turn affects the stability of LC. In space, closing the left lane, the working area of the middle lane and the diversion area of the interchange have a great influence on the LC pattern, showing more lateral rapid acceleration patterns and deceleration patterns. These findings have certain reference significance for road geometric design. The heterogeneity of LC trajectory patterns under different road geometries suggests that the data in different road environments should be used to train the best parameters of the identification method.

Full time domain trajectory big data has the advantages of low cost, large amount of information, wide coverage and high frequency, and is a good source of traffic information collection data. It provides a ready-made solution for collecting driving data for LC pattern recognition and further related research. We propose a method to prepare full-time domain LC data before realizing LC recognition, including road fitting, point-line matching and LC rule judgment. The comparison of the results shows the advantages of our method framework and further enriches the application of high-frequency GPS data. The clustering method of change point detection and similarity matching is applied to identify LCs in full-time domain vehicle trajectory data, which can be used as a new data mining solution and provide ideas for traffic design in complex scenes. But in the end, it is worth pointing out the limitations of our research. Although our data has a wide range of characteristics, it is difficult to collect the information of vehicles around LC. Although this is not important to the theme of this paper about the background of big data in full-time domain, considering it in the future can improve the accuracy of the method. In addition, although the trajectory of our data is a heavy truck, which can represent the most unfavorable LC influence in the expressway scene, it lacks a comparison of the LC behavior of small cars and the heterogeneity of LC among different drivers. In the future, we will overcome these limitations and consider the specific LC behavior in specific scenarios, such as the mandatory LC behavior in the junction area of interchange, to supplement our application. In addition, in the future era of self-driving cars, through the method proposed in this paper to explore the optimal design under the condition of self-driving, the risk LC pattern of self-driving cars in these areas can be reduced, the accident risk can be reduced and the traffic smoothness can be improved.

References

Ali, Yasir, Michiel C. J. Bliemer, Zuduo Zheng, and Md Mazharul Haque. 2020. “Cooperate or Not? Exploring Drivers’ Interactions and Response Times to a Lane-Changing Request in a Connected Environment.” TRANSPORTATION RESEARCH PART C-EMERGING TECHNOLOGIES 120.

Ali, Yasir, Zuduo Zheng, Md Mazharul Haque, Mehmet Yildirimoglu, and Simon Washington. 2020. “Detecting, Analysing, and Modelling Failed Lane-Changing Attempts in Traditional and Connected Environments.” ANALYTIC METHODS IN ACCIDENT RESEARCH 28.

Ali, Yasir, Zuduo Zheng, Md. Mazharul Haque, Mehmet Yildirimoglu, and Simon Washington. 2021. “CLACD: A Complete LAne-Changing Decision Modeling Framework for the Connected and Traditional Environments.” TRANSPORTATION RESEARCH PART C-EMERGING TECHNOLOGIES 128.

Chen, Hailan, and Xuedong Gao. 2020. “A New Time Series Similarity Measurement Method Based on Fluctuation Features.” TEHNICKI VJESNIK-TECHNICAL GAZETTE 27(4):1134–41.

Chen, Qinghong, Helai Huang, Ye Li, Jaeyoung Lee, Kejun Long, Ruifeng Gu, and Xiaoqi Zhai. 2021. “Modeling Accident Risks in Different Lane-Changing Behavioral Patterns.” ANALYTIC METHODS IN ACCIDENT RESEARCH 30.

Chen, Shuyan, Hong Yao, Fengxiang Qiao, Yongfeng Ma, Ying Wu, and Jian Lu. 2023. “Vehicles Driving Behavior Recognition Based on Transfer Learning.” EXPERT SYSTEMS WITH APPLICATIONS 213(C).

Chen, Yaoyu, Guofa Li, Shen Li, Wenjun Wang, Shengbo Eben Li, and Bo Cheng. 2022. “Exploring Behavioral Patterns of Lane Change Maneuvers for Human-Like Autonomous Driving.” IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS 23(9):14322–35.

Cheng, Bo, Da Zhu, Shuai Zhao, and Junliang Chen. 2016. “Situation-Aware IoT Service Coordination Using the Event-Driven SOA Paradigm.” IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT 13(2):349–61.

De Almeida, Raul Tomaz, Antonio Luis Pimentel Vasconcelos, and Ana Maria Cesar Bastos Silva. 2018. “DESIGN CONSISTENCY INDEX FOR TWO-LANE ROADS BASED ON CONTINUOUS SPEED PROFILES.” PROMET-TRAFFIC & TRANSPORTATION 30(2):231–39.

Di, Yangchen, Mingyue Lu, Min Chen, Zhangjian Chen, Zaiyang Ma, and Manzhu Yu. 2022. “A Quantitative Method for the Similarity Assessment of Typhoon Tracks.” NATURAL HAZARDS 112(1):587–602.

Elspas, Philip, Yannick Klose, Simon Isele, Johannes Bach, and Eric Sax. 2021. “Time Series Segmentation for Driving Scenario Detection with Fully Convolutional Networks” edited by K. Berns, M. Helfert, and O. Gusikhin. PROCEEDINGS OF THE 7TH INTERNATIONAL CONFERENCE ON VEHICLE TECHNOLOGY AND INTELLIGENT TRANSPORT SYSTEMS (VEHITS) 56–64.

Fan, Pengcheng, Jingqiu Guo, Yibing Wang, and Jasper S. Wijnands. 2022. “A Hybrid Deep Learning Approach for Driver Anomalous Lane Changing Identification.” ACCIDENT ANALYSIS AND PREVENTION 171.

Guijo-Rubio, David, Antonio Manuel Duran-Rosal, Pedro Antonio Gutierrez, Alicia Troncoso, and Cesar Hervas-Martinez. 2021. “Time-Series Clustering Based on the Characterization of Segment Typologies.” IEEE TRANSACTIONS ON CYBERNETICS 51(11):5409–22.

Hamedi, Hamidreza, and Rouzbeh Shad. 2022. “Context-Aware Similarity Measurement of Lane-Changing Trajectories.” EXPERT SYSTEMS WITH APPLICATIONS 209.

Hamedi, Hamidreza, Rouzbeh Shad, and Sadegh Jamali. 2023. “Measuring Lane-Changing Trajectories by Employing Context-Based Modified Dynamic Time Warping.” EXPERT SYSTEMS WITH APPLICATIONS 216.

Hamedi, Hamidreza, Rouzbeh Shad, and Seyed Ali Ziaee. 2022. “A Comparative Study on Measurement of Lane-Changing Trajectory Similarities.” PHYSICA A-STATISTICAL MECHANICS AND ITS APPLICATIONS 604.

Haynes, Kaylea, Paul Fearnhead, and Idris A. Eckley. 2017. “A Computationally Efficient Nonparametric Approach for Changepoint Detection.” STATISTICS AND COMPUTING 27(5):1293–1305.

Jokhio, Sarang, Pierluigi Olleja, Jonas Baergman, Fei Yan, and Martin Baumann. 2023. “Analysis of Time-to-Lane-Change-Initiation Using Realistic Driving Data.” IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

Li, Min, Yumei Huang, and Youwei Wen. 2023. “A Total Variation Based Method for Multivariate Time Series Segmentation.” ADVANCES IN APPLIED MATHEMATICS AND MECHANICS 15(2):300–321.

Long, Yan, Jianling Huang, Xiaohua Zhao, and Zhenlong Li. 2022. “Does LSTM Outperform 4DDTW-KNN in Lane Change Identification Based on Eye Gaze Data?” TRANSPORTATION RESEARCH PART C-EMERGING TECHNOLOGIES 137.

Panek, Michal, Ireneusz Jablonski, and Michal Wozniak. 2022. “Automated Identification of Systematic Performance Changes in 5G Networks by Changepoint Tracking.” 2022 INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN).

Panichpapiboon, Sooksan, and Puttipong Leakkaw. 2020. “Lane Change Detection With Smartphones: A Steering Wheel-Based Approach.” IEEE ACCESS 8:91076–88.

Qi, Jin Peng, Fang Pu, Ying Zhu, and Ping Zhang. 2022. “A Weighted Error Distance Metrics (WEDM) for Performance Evaluation on Multiple Change-Point (MCP) Detection in Synthetic Time Series.” COMPUTATIONAL INTELLIGENCE AND NEUROSCIENCE 2022.

Qu, Dayi, Kekun Zhang, Hui Song, Tao Wang, and Shouchen Dai. 2022. “Analysis of Lane-Changing Decision-Making Behavior of Autonomous Vehicles Based on Molecular Dynamics.” SENSORS 22(20).

Shi, Ruian, and Quaid Morris. 2021. “Segmenting Hybrid Trajectories Using Latent ODEs” edited by M. Meila and T. Zhang. INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 139 139.

Tselentis, Dimitrios I., and Eleonora Papadimitriou. 2023. "Driver Profile and Driving Pattern Recognition for Road Safety Assessment: Main Challenges and Future Directions." IEEE OPEN JOURNAL OF INTELLIGENT TRANSPORTATION SYSTEMS 4:83–100.

Wang, Jing, and ZhongCheng Wu. 2023. "Driver Distraction Detection via Multi-Scale Domain Adaptation Network." IET INTELLIGENT TRANSPORT SYSTEMS 17(9):1742–51.

Wu, Jian-Da, and Chia-Hsin Chang. 2022. "Driver Drowsiness Detection and Alert System Development Using Object Detection." TRAITEMENT DU SIGNAL 39(2):493–99.

Zhang, Yuan-qiang, and Guo-you Shi. 2021. "Trajectory Similarity Measure Design for Ship Trajectory Clustering." 2021 IEEE 6TH INTERNATIONAL CONFERENCE ON BIG DATA ANALYTICS (ICBDA 2021) 181–87.

Zhang, Yue, Yajie Zou, Yunlong Selpi, Yunlong Zhang, and Lingtao Wu. 2023. "Spatiotemporal Interaction Pattern Recognition and Risk Evolution Analysis During Lane Changes." IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS 24(6):6663–73.

Zheng, Yandong, Rongxing Lu, Yunguo Guan, Jun Shao, and Hui Zhu. 2022. "Efficient and Privacy-Preserving Similarity Range Query Over Encrypted Time Series Data." IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING 19(4):2501–16.

Zheng, Zuduo, and Simon Washington. 2012. "On Selecting an Optimal Wavelet for Detecting Singularities in Traffic and Vehicular Data." Transportation Research Part C: Emerging Technologies 25:18–33. doi: https://doi.org/10.1016/j.trc.2012.03.006.

Acknowledgments

The authors would like to thank the National Key Research and Development Program of China (No. 2020YFC1512002), Sichuan Science and Technology Program (No. 2022YFG0048], Science and Technology Project of Sichuan Transportation Department (No. 2022-ZL-04), and Key Research and Development Program of Shanxi Province (No. 202102020101014). Their assistance is gratefully acknowledged.

Min Zhang is an associate professor and holds a PhD in Engineering in Transportation Planning and Management from Chang'an University. My research areas include traffic safety theory and technology, traffic planning and design, comprehensive transportation system analysis, etc. I have led or participated in national key research and development programs, central university high-tech research and cultivation projects, World Bank urban transportation cooperation projects (Xi'an Public Transport Operation Improvement Research), Xi'an urban transportation development research, and other scientific research projects. Published over 20 academic papers, won 1 third prize in science and technology from the China Highway Society, authorized 2 national invention patents, 1 utility model patent, and 3 software copyrights.

Yuhan Nie graduated from Changsha University of Science and Technology with a major in Transportation Engineering, and began pursuing a doctoral degree in Transportation Engineering at Chang'an University in 2024. He mainly engages in the development of traffic engineering software, as well as research on traffic big data technology and traffic safety.

Chi Zhang, Ph.D. in Engineering, postdoctoral experience, professor at Chang'an University, recipient of the Wu Fu Zhenhua Transportation Education Outstanding Teacher Award, doctoral/international student supervisor, postdoctoral collaborative supervisor, visiting scholar at Clemson University in the United States, registered road engineer, engaged in research and teaching work in road overall design, road digitization, traffic safety, and interdisciplinary fields. Hosted and participated in over 40 scientific research projects, including national key research and development programs. Edited and collaborated on the publication of 2 textbooks and academic monographs. Published over 100 academic papers as first author/corresponding author, including more than 50 SCI/EI indexed papers and 10 authorized invention patents. Received 8 national and provincial-level teaching achievement awards, 9 provincial-level scientific research awards or honors, and guided students to receive 15 national and provincial-level awards.