Journal of Engineering, Project, and Production Management
2026, 16(5), 2025-359

A Hybrid K-Means++ and Gaussian Mixture Model for High-Accuracy Travel Behavior Prediction Using ETC Data

Dandan Wang
Lecturer, School of Artificial Intelligence, Zibo Polytechnic University, Zibo, 255000, China, E-mail: DandanWangzz@outlook.com

Project Management
Received December 25, 2025; revised February 10, 2026; accepted July 7, 2026
Available online July 20, 2026

Abstract: With the continuous expansion of highway networks and the in-depth development of intelligent transportation systems, the Electronic Toll Collection (ETC) covers a large amount of user behavior data information. To enhance the accuracy of user travel prediction, this study proposes a hybrid clustering framework that employs the K-means++ algorithm to conduct efficient preliminary clustering on user travel behavior data. After determining the initial parameter range, the Gaussian Mixture Model (GMM) is used for refined iterative optimization. The accuracy in user travel prediction reached 94.83%, the recall was 93.87%, the F1 score was 94.25%, and the prediction latency was only 45.53ms, supporting 950 concurrent requests per second. In addition, on the clustering index, the silhouette coefficient of the hybrid model reached 0.64, the Calinski-Harabasz Index (CHI) was 5,861, and the Davies-Bouldin Index (DBI) was 0.68. In summary, the prediction model effectively improves the accuracy of user travel predictions and the efficiency of practical applications, and provides support and a reference for transportation network planning and personalized travel recommendations.

Keywords: Electronic toll collection (ETC); k-means++; gaussian mixture model (GMM) hybrid clustering; big data mining; traffic behavior pattern

DOI 10.32738/JEPPM-2025-359

1. Introduction

With the rapid development of highway networks, transportation efficiency has significantly improved, greatly promoting social and economic development. The continuous advancement of information technology has made intelligent transportation systems widely used in major high-speed networks. Intelligent transportation systems have become a key part of effective traffic management because they can improve traffic efficiency, enhance service quality, and monitor road traffic information in real time (Oatley 2022). The Electronic Toll Collection (ETC) gantry system, as a key infrastructure of the intelligent transportation system, has accumulated a massive amount of user travel data. These data include important information, such as users' spatiotemporal travel patterns, route-selection preferences, and traffic distribution characteristics (Shahrier et al., 2024). However, user travel behavior is affected by many factors, such as weather, holidays, traffic events, and personal habits, and exhibits significant nonlinear, dynamic, and uncertain characteristics (Chango et al. 2022). Traditional travel prediction methods are mostly based on statistical methods like regression analysis and time series analysis, which are difficult to effectively capture complex user behavior patterns. Therefore, there are shortcomings in predictive accuracy and generalization. K-means++ Clustering (K-means++) is often applied in data mining because it can improve clustering quality and stability while reducing dependence on randomness (Sarwar et al., 2022). The Gaussian Mixture Model (GMM) can cluster samples from a probability perspective and can handle the time discontinuity in ETC data (Yu et al., 2022). Therefore, in this context, the study proposes a method based on K-means++ and GMM, which uses K-means++ for preliminary clustering to quickly determine an approximate classification of user behavior and reduce sensitivity to initial values, and then GMM iterates to hierarchically cluster ETC data. The Entropy Weight Method (EWM) is also introduced to optimize the weights of user travel characteristics to more effectively mine and analyze ETC data, capture complex user behavior patterns, and improve the accuracy and reliability of user travel predictions in ETC big data.

The core contributions of the research are mainly reflected in three aspects: Firstly, proposing a hybrid K-means++ and GMM clustering model that integrates EWM, objectively optimizing user travel feature weights through EWM, and solving the traditional clustering algorithms being sensitive to initial values and not performing well in processing mixed density data; Secondly, constructing an ETC user travel prediction architecture based on Multi-modal Spatiotemporal Sequence (MSTSP), integrating multi-dimensional features of space, time, and users, and combining multi-head attention mechanism

and Graph Convolutional Network (GCN) to accurately capture the nonlinear spatiotemporal dependencies of ETC data; Thirdly, empirical verification was conducted based on large-scale ETC transaction data from real provinces, which not only completed multidimensional evaluations of clustering effectiveness and prediction performance, but also tested the deployment indicators, including prediction latency, concurrent processing capability, and resource utilization, providing a quantitative basis for the actual implementation of the model.

2. Related Works

2.1. ETC Data Mining and Toll Trajectory Prediction

With the expansion of the highway network and the development of intelligent transportation technology, the massive transaction data accumulated by the ETC gantry system has become the core data source for analyzing user travel behavior. Domestic and foreign scholars have conducted extensive research on ETC data mining and toll trajectory prediction. To accurately identify the risk factors of traffic crashes and improve highway safety, Yang et al. (2022) built a big data mining method based on weighted orientation. This method could better reveal the causes of highway traffic collapse and make more accurate traffic predictions. To prevent traffic accidents, Lakshmi et al. (2022) proposed a traffic response system based on the Internet of Things (IoT). By analyzing transportation data, unexpected traffic stop events could be effectively monitored and managed, thereby significantly reducing the accident rate on fast-moving roads and highways. These studies indicate that ETC data mining is a key foundation for improving travel prediction accuracy, but existing methods still have limitations in handling data nonlinearity and dynamism.

2.2. Clustering for Mobility Pattern Mining

The clustering algorithm is the core technology for mining user travel patterns, with K-means++ and GMM widely used in transportation. Zhang and Sun (2024) proposed an intelligent travel planning system based on kernel density estimation and K-means++. The system could perform high-precision travel prediction. Park et al. (2022) developed a travel speed prediction model using real-time data. This model used K-means++ to predict travel speed. The predicted travel speed data generated by this model helped drivers choose optimized paths. Chen et al. (2023) built a Bayesian probability model to predict bus travel time probabilistically and estimate arrival time. This method used a GMM to describe the distribution of bus travel time through constrained multivariate Gaussian distributions. The method was effective. To optimize the accuracy of path travel-time estimation in urban areas, Gemma et al. (2024) developed a GMM-based method. The method exploited heterogeneous travel-time measurement data from different monitoring technologies. This method could significantly improve the accuracy of travel time estimation. The above research has verified the effectiveness of clustering algorithms in travel pattern mining, but each clustering algorithm has challenges, such as sensitivity to initial values and limited ability to handle mixed-density data.

2.3. Spatiotemporal Deep Learning for Travel Forecasting

The spatiotemporal deep learning method has become a research hotspot in travel prediction in recent years due to its powerful nonlinear fitting capabilities. Long short-term memory networks, Convolutional Neural Networks (CNNs), GCNs, and attention mechanisms are widely used to capture spatiotemporal dependencies in travel data. Lai et al. (2022) proposed an Internet-based ETC big data mining method that combines two-way short-term memory networks, conditional random fields and Multilayer Perceptions (MLPs) to achieve deep data analysis. The results show that bidirectional long short-term memory networks can effectively handle the long-term dependency challenge in time-series data and achieve good performance in predicting ETC toll trajectories. Zhang et al. (2023) proposed a CNN-based big data mining method to improve the efficiency of information collection and extraction in current intelligent transportation systems. The results show that CNN can extract spatial features from vehicle trajectory data through its local feature-extraction capability, resulting in a prediction accuracy of 75.2%. GCN can model spatial correlations among road segments based on the road network's topology, thereby improving the accuracy of regional traffic flow prediction. The multi-head attention mechanism can adaptively focus on key spatiotemporal features and reduce interference from redundant information in prediction results. However, most existing spatiotemporal deep learning methods have high computational complexity and are difficult to meet the real-time prediction needs of ETC systems.

In summary, although research on traffic flow analysis, such as big data mining, has achieved many results, hybrid K-means++ and GMM clustering still have some limitations. For example, most methods have high computational complexity and are difficult to apply in real-time. Therefore, this study combines the advantages of both methods and proposes a hybrid clustering model to optimize prediction accuracy and real-time performance, thereby satisfying actual needs.

3. User Travel Prediction Model Based on Hybrid Clustering

A hybrid clustering model is built using K-means++ and GMM to perform pre-clustering and fine-grained clustering. Then, the EWM is used to optimize feature weights, improving the accuracy and stability of the clustering results.

3.1. Hybrid Clustering Model Based on K-Means++ and GMM

With the continuous expansion of highway networks and the widespread application of intelligent transportation systems, massive ETC user travel data provides a rich foundation for analyzing user behavior patterns. However, these travel data are nonlinear, dynamic and uncertain, making it hard for traditional statistical methods to capture complex user behavior patterns (Gopinathan et al., 2024). K-means++ is widely used in various data mining tasks for its ability to significantly improve clustering quality and stability and reduce sensitivity to initial values (Xie et al., 2023). GMM can model data through a probability distribution, effectively handling the time discontinuity challenge of data (Saygılı 2022). Therefore, the study combines the advantages of the above two clustering algorithms to design a hybrid clustering model, as shown in Fig. 1.

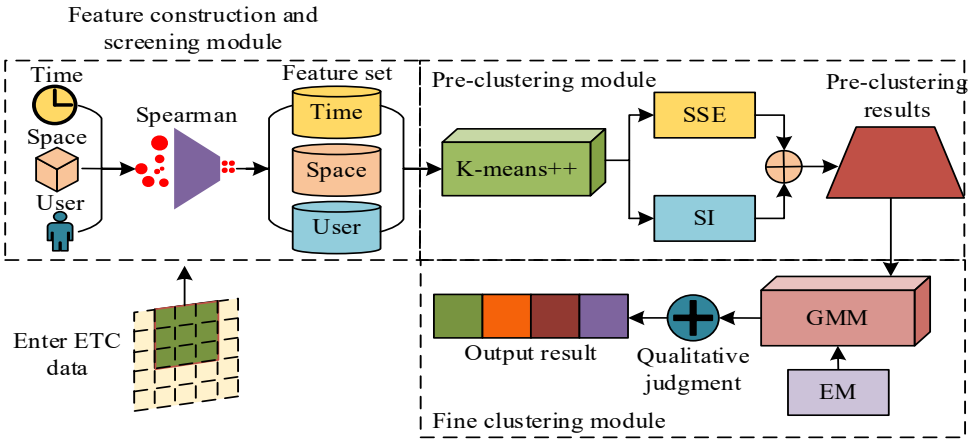


Fig. 1. A hybrid clustering model based on K-means++and GMM

In Fig. 1, the structure mainly has feature construction and screening modules, pre-clustering modules, and precise clustering modules. In the feature construction and screening module, the ETC data is filtered and divided into three attributes, yielding three types of feature sets. After preliminary clustering, the optimal K value is judged through indicators to obtain the pre-clustering result. Then, the GMM is used for fine clustering, and the Expectation-Maximization (EM) estimates parameters to further optimize the pre-clustering results. Finally, a qualitative judgment is made on the clustering results to identify the user's main travel mode. The Spearman correlation coefficient can capture nonlinear relationships in the data by quantifying the correlation between variables, as shown in Eq. (1).

$$\rho = 1 - \frac{6\sum \alpha_a^2}{m\left(m^2 - 1\right)} \tag{1}$$

In Eq. (1), ρ represents the Spearman correlation coefficient. α_a represents the difference between the rankings of the a-th observation in the two variables. m signifies the number of samples in the data set. The study uses SSE to measure the dispersion of clustering results, as shown in Eq. (2).

$$SSE = \sum_{i=1}^{I} \sum_{p \in C_i} |x_i - c_i|^2 \tag{2}$$

In Eq. (2), C_i signifies the cluster of the i-th cluster. x_i represents the i-th point in the cluster. c_i signifies the cluster center point. The SI is shown in Eq. (3).

$$SI = \sum_{i=1}^{I} \frac{\bar{B}_i - \bar{A}_i}{max(\bar{A}_i, \bar{B}_i)} \tag{3}$$

In Eq. (3), $\bar{A}_i$ signifies the average distance from i to other points in the same cluster. $\bar{B}_i$ stands for the average distance from i to all points in the nearest other clusters. The Euclidean distance of K-means++ clustering is shown in Eq. (4).

$$D(x_i, c_i) = \sqrt{\sum_{t=1}^{T} (x_{it} - c_{it})^2} \tag{4}$$

In Eq. (4), D stands for the Euclidean distance. T signifies the total number of dimensions. x_{it} signifies the value of point x_i in the t-th dimension. c_{it} signifies the value of the cluster center point c_i in the t-th dimension. GMM assumes that data points come from multiple Gaussian distributions, and the probability density function is presented in Eq. (5).

$$P(x) = \sum_{k=1}^{K} \omega_k \cdot \Gamma(x|\mu_k, \Phi_k) \tag{5}$$

In Eq. (5), $P(x)$ stands for the probability density function of data point x. ω_k signifies the weight of the k-th Gaussian distribution. K signifies the total number of Gaussian distributions. $\Gamma(x|\mu_k, \Phi_k)$ signifies the k-th Gaussian distribution with mean μ_k and covariance matrix Φ_k. ETC data not only contains users' travel characteristics, but also contains many redundant feature information. The EWM can determine the weight by calculating the information entropy of each indicator, reflecting the importance of each indicator more objectively (Dou et al., 2023). Therefore, to further optimize the feature construction and screening process and improve the accuracy and stability of clustering results, the EWM is employed to optimize the hybrid clustering process of K-means++ and GMM. The optimized hybrid clustering process based on K-means++ and GMM is shown in Fig. 2.

In Fig. 2, in this process, data sets of different feature types are first constructed based on data features. Then, the EWM calculates the information entropy of each feature. The weight of each feature is calculated based on the information entropy. Next, the K-means++ is used to initialize the cluster centers and assign the features to the nearest class. Based on K-means++, the mean vector, covariance matrix and mixing coefficient are initialized. At this time, whether the maximum number of iterations has been reached is determined. If the maximum number of iterations is met, the final result is output. If the maximum number of iterations is not reached, the EM algorithm is used to iteratively calculate the parameters and return to the initialized mean vector until the termination condition is met. The study first standardizes the indicators of each feature, as shown in Eq. (6).

$$y_{ij}^{'} = \frac{y_{ij} - min(y_{ij})}{max(y_{ij}) - min(y_{ij})} \tag{6}$$

In Eq. (6), $y_{ij}^{'}$ signifies the standardized result of the i-th feature on the j-th index. y_{ij} represents the index before normalization. $max(y_{ij})$ and $min(y_{ij})$ signify the maximum and minimum values of the j-th indicator among all features. The proportion of indicators is shown in Eq. (7).

$$p_{ij} = \frac{y_{ij}^{'}}{\sum_{i=1}^{J} y_{ij}^{'}} \tag{7}$$

In Eq. (7), p_{ij} refers to the proportion of the i-th feature on the j-th indicator. J indicates the total number of indicators. The information entropy value is shown in Eq. (8).

$$E_j = -\frac{\sum_{i=1}^{J} p_{ij} \ln p_{ij}}{\ln J} \tag{8}$$

In Eq. (8), E_j signifies the information entropy of the j. The indicator weight is shown in Eq. (9).

$$w_j = -\frac{1 - E_j}{\sum_{i=1}^{J} (1 - E_j)} \tag{9}$$

In Eq. (9), w_j signifies the weight of the j-th indicator.

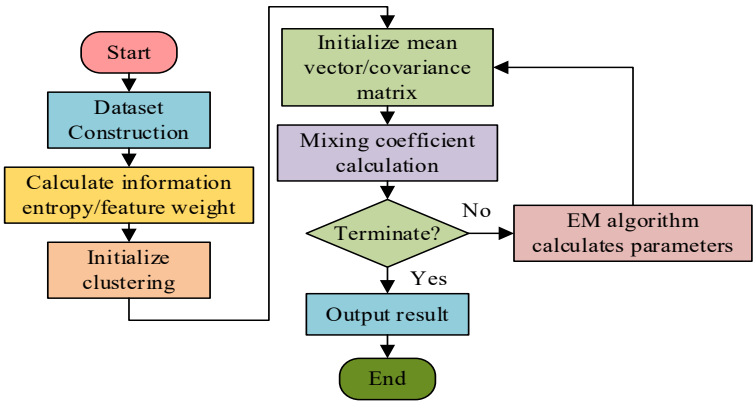


Fig. 2. Optimized hybrid clustering process based on K-means++ and GMM

3.2. User Travel Prediction Model Based on Multi-Modal Spatiotemporal Sequence

The ETC data exhibit complex nonlinear spatiotemporal dependencies, posing challenges for extracting and predicting user travel features. Therefore, after using the hybrid clustering model to accurately extract ETC feature data, to further achieve accurate user travel prediction, it is necessary to fully capture the complex spatiotemporal dependencies in the user's travel process. The study introduces the MLP and constructs a user travel prediction model based on MSTSP, as presented in Fig. 3.

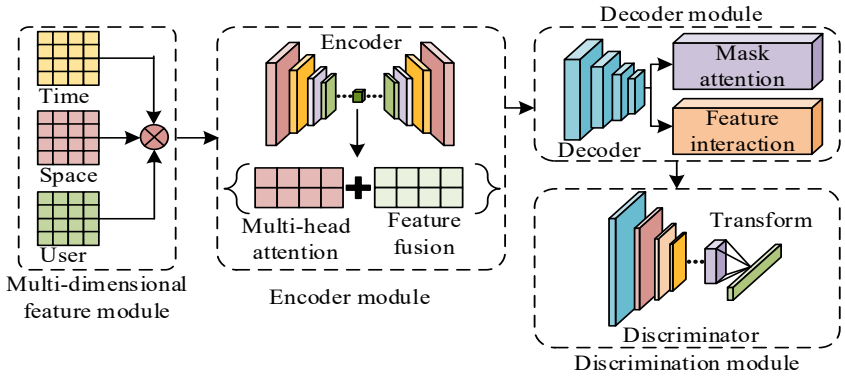


Fig. 3. User travel prediction model based on MSTSP

From Fig. 3, the model mainly includes a multi-dimensional feature module, an encoder module, a decoder module and a discriminant module. The multi-dimensional feature module covers features across space, time and users to uniformly vectorize discrete and continuous features. The multi-head attention mechanism in the encoder module can focus on key feature interactions through sparse attention weights, thereby capturing long-range dependencies among multi-dimensional features. The feature fusion module uses feature splicing operations to fuse features into unified representations that capture global spatiotemporal information. The mask attention module in the decoder ensures causal prediction through masking operations, while the feature interaction module gradually generates a preliminary prediction feature sequence after receiving

the global features produced by the encoder. As the core structure of MLP, the discriminant module performs a nonlinear transformation and error correction on the preliminary prediction features output by the decoder. Finally, it maps the high-dimensional features to the target prediction space via a linear transformation and outputs the predictions. The sequence length and prediction duration of the MSTSP model are set as follows: The study sets the input spatiotemporal sequence length to 24, corresponding to 24-hour historical travel data. The predicted duration can be flexibly configured to range from 1 to 6 hours, meeting the needs of short-term travel prediction. When constructing the sequence, a sliding time window is used to generate the input sequence and corresponding predicted labels by rolling forward in 1-hour increments. The encoder is presented in Eq. (10).

$$X_{b+1}^{t} = MaxPool\left(GELU\left(Convld\left(\left[X_{b}^{t}\right]_{A}\right)\right)\right) \tag{10}$$

In Eq. (10), X_{b+1}^{t} signifies the output of the $b+1$-th layer encoder at time step t. X_{b}^{t} signifies the input of the b-th layer at time step t. $[\cdot]_{A}$ signifies the attention mechanism. $Convld$ represents a one-dimensional convolution operation. $GELU$ means applying a Gaussian error linear unit activation function to the convolution result. $MaxPool$ means applying max pooling operation to the $GELU$ activation result. The feature fusion is shown in Eq. (11).

$$X_{feet}^{t} = w_1 \times S_c + w_2 \times X_c^{t} \tag{11}$$

In Eq. (11), X_{feet}^{t} signifies the features after fusion at time step t. w_1 and w_2 represent the influence weight of spatial features and time features, respectively. S_c represents the spatial characteristics of node c. X_c^{t} represents the time characteristics of node c at time step t (Chen 2022). The attention mechanism is shown in Eq. (12).

$$A(Q,K,V) = Soft\,max\left(\frac{QK^{T}}{\sqrt{d}}\right)V \tag{12}$$

In Eq. (12), A represents the attention mechanism. Q, K and V refer to query, key and value matrices. T means transpose. d signifies the dimension of the key vector. To accurately predict the user's travel destination, the physical connectivity and proximity of the user's desired travel interval and the functional similarity between locations are analyzed in depth. Therefore, when embedding spatial-dimensional features, the dynamic attention mechanism is used to process multiple spatial matrices and adapt to time-varying data characteristics. The multiple-spatial-attention framework is shown in Fig. 4.

In Fig. 4, this framework uses distance matrices, proximity matrices, and functional similarity matrices as inputs for spatial features to perform dynamic attention calculations. Through the graph-structured data modeling capability of GCN, the spatial topological relationships and nonlinear characteristics implicit in various matrices can be mined to map the original data to a high-dimensional feature space. Finally, dimension fusion is performed via matrix splicing to form a comprehensive feature matrix that incorporates multi-dimensional spatial features. The study uses the Haversine Formula (HF) to calculate the shortest distance between two points on the sphere. The distance matrix is presented in Eq. (13).

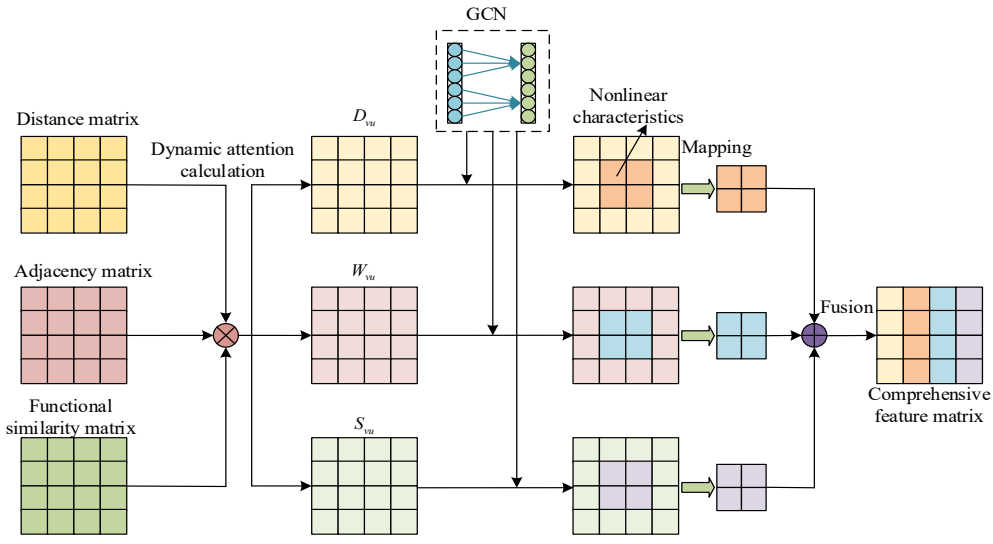


Fig. 4. Multi-spatial attention framework

$$D_{vu} = R \times 2\,arcsin\left(\sqrt{sin^2\left(\frac{\phi_2-\phi_1}{2}\right) + cos(\phi_1) \times cos(\phi_2) \times sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right)}\right) \tag{13}$$

In Eq. (13), D_{vu} represents the distance between points v and u. R represents the radius of the earth. ϕ_1 and ϕ_2 signify the latitude of points v and u. φ_1 and φ_2 signify the longitude of points v and u. The proximity matrix is presented in Eq. (14).

$$W_{vu} = \frac{\omega(v,u)}{\sum_o \omega(v,o)} \tag{14}$$

In Eq. (14), W_{vu} represents the normalized proximity matrix. $\omega(v,u)$ represents the connection weight between points v and u in the adjacency matrix. $\sum_o \omega(v,o)$ represents the sum of all neighbor nodes o of point v. The functional similarity matrix is shown in Eq. (15).

$$S_{vu} = \frac{E(D_v', D_u') - E(D_v')E(D_u')}{\sigma(D_v') \times \sigma(D_u')} \tag{15}$$

In Eq. (15), S_{vu} represents the functional similarity matrix. D_v' and D_u' represent the functional feature vectors of points v and u, respectively. $E(D_v', D_u')$ represents the joint expected value of D_v' and D_u'. $\sigma(D_v')$ and $\sigma(D_u')$ represent the standard deviations of D_v' and D_u', respectively. The MSTSP model leverages the feature outputs of hybrid clustering and accurately captures nonlinear spatiotemporal correlations in ETC data through multi-module collaboration and a GCN-enhanced dynamic attention mechanism, thereby providing reliable support for user travel prediction. The construction method and update mechanism of the spatial matrix are as follows: the distance matrix, proximity matrix, and functional similarity matrix involved in the study are all constructed based on the attribute information and topological relationship of ETC gantry nodes. The distance matrix calculates the shortest spherical distance between the gantry nodes using the Haversine formula in Eq. (13). The adjacency matrix is constructed based on the topological relationship of the highway network. If two gantry nodes are directly connected, the weight is assigned to 1. Otherwise, it is set to 0, and normalization is performed using Eq. (14). The functional similarity matrix is constructed based on the functional attributes of the road section to which the gantry node belongs. The Pearson correlation coefficient of the functional feature vectors of different nodes is calculated using Eq. (15) to quantify the functional similarity between nodes. The update of the spatial matrix uses a periodic rolling strategy, adjusting weekly data changes based on road network topology and functional attributes, and gradually updating the matrix to ensure it adapts to the road network's dynamic changes. In the multi-space attention framework, the GCN models distance, proximity, and functional similarity matrices separately. For each spatial matrix, an independent two-layer GCN structure is constructed, and a graph convolution operation is used to capture the spatial topological relationships and nonlinear features within each matrix, thereby producing high-dimensional spatial feature vectors for the corresponding matrix. In the feature fusion stage, attention weighted fusion strategy is adopted to calculate the attention weight of each matrix feature vector. The weight size is determined by the correlation between the feature vector and the global spatiotemporal features. Finally, the fused comprehensive spatial feature matrix is obtained by weighted summation, achieving complementary and enhanced multidimensional spatial information.

4. Validation of User Travel Prediction Model Based on Hybrid Clustering

4.1. Experimental Environment Setup

To validate the effectiveness of the proposed user travel prediction model based on hybrid clustering, experiments were conducted on Ubuntu 16.04.7 LTS. The Adam optimization algorithm is selected for model training, with a learning rate of 0.0001. The number of heads in the multi-head attention mechanism is 8, and the number of GCN layers is 2. The batch size is 256 and the training epoch is 300. The CUDA toolkit version in the experimental environment is 10.1, which is fully compatible with cuDNN 7.6.5, and together they provide stable support for GPU-accelerated computing. PyTorch 1.0 needs to be compiled and installed with CUDA 10.1 to ensure compatibility with the underlying acceleration libraries. NumPy version 1.19.5 and Matplotlib version 3.3.4 have no version conflicts with Python 3.8.0 and PyTorch 1.0. The data set comes from actual ETC transaction data provided by a highway network toll center in a certain province in China. The time span is from January 2023 to June 2024. It contains a total of about 15 million active ETC user cards, and about 850 million pieces of transaction data are recorded. After the data is cleaned and preprocessed, 70% is used for training, 15% is used for validation, and 15% is used for testing, of which the validation set is used for hyperparameter tuning. Table 1 presents the experimental environment configuration.

Table 1. Experimental environment configuration

Environment	Configuration
Operating System	Ubuntu 16.04.7 LTS
GPU	NVIDIA GeForce GTX 1080Ti
CPU	Intel(R) Core(TM) i5-6500 @ 3.20GHz
Memory	32GB
Programming Language	Python 3.8.0
Development Framework	PyTorch 1.0
Data Processing Library	NumPy
Visualization Library	Matplotlib
Accelerated Computing Library	cuDNN 7.6.5

4.2. Clustering Effect Verification

To verify the clustering effect, the study conducted a visual analysis of the results from separate K-means++ and GMM runs, as shown in Fig. 5. From Fig. 5(a), the original data distribution contained points of four different colors. In the K-means++ clustering result in Fig. 5(b), the data of different colors were roughly clustered. However, the red and blue clusters in the lower-left corner were mistakenly merged, while the high-density area in the upper-right corner was split into two clusters, indicating insufficient segmentation accuracy. In the GMM clustering result in Fig. 5(c), the outlier point in the upper left

corner formed an invalid small cluster. Although GMM achieves better clustering within the same density region, its adaptability is limited when dealing with hybrid density regions. The hybrid clustering model results in Fig. 5(d) showed that the boundaries between clusters were clear and there were no invalid clusters. This method effectively avoided interference from outliers and accurately separated hybrid density regions. In summary, the hybrid clustering model based on K-means++ and GMM performs well with complex data distributions, effectively addresses challenges such as uneven density and overlapping clusters, and achieves excellent clustering results.

To further validate the clustering performance in the actual ETC feature space, the study selected four core feature dimensions of ETC data, namely travel time period, mileage, road segment type preference, and weekly travel frequency, and conducted quantitative statistical analysis on the clustering results of the optimal K value. The results are shown in Table 2. The proportions of morning rush-hour and weekly high-frequency trips for commuting both exceeded 65% and were mainly for short-distance and urban high-speed travel. The weekly travel frequency was high, reflecting urban residents who commuted daily with fixed routes. The proportion of long-distance business travel exceeded 65%, with a preference for inter-provincial highways accounting for over 70%. The frequency of weekly travel was moderate, and the corresponding group comprised cross-city business travelers, with work needs as the primary travel purpose. The preference for suburban highways for holiday sunrise travel was obvious, with long-distance travel accounting for about 45%. The weekly travel frequency was low, and the corresponding group consisted of weekend or holiday travel users. The travel frequency has periodicity. The proportion of occasional travel across different time periods was uniform; the mileage distribution was random; the weekly travel frequency was extremely low; and the corresponding group consisted of temporary travel users with no fixed travel pattern. Based on the above, the hybrid clustering model can effectively distinguish the travel patterns of different users within the actual ETC feature space, and the clustering results have clear business value.

Table 2. Quantitative evaluation results of clustering based on actual ETC feature space

Data type	Average proportion of morning rush hour/%	Average proportion of long-distance travel/%	Proportion of preference for inter provincial highways/%	Proportion of high-frequency travel on a weekly basis/%
Commuting type travel	68.20	12.53	8.34	75.13
	72.12	9.84	10.26	80.58
Long distance business travel	15.33	65.79	72.44	42.85
	18.54	70.27	78.10	38.64
Holiday sunrise tour type travel	22.43	45.33	25.76	15.25
	20.13	42.81	28.36	18.73
Occasional travel style	8.64	20.57	15.46	5.34
	10.20	18.70	12.63	4.87

To verify the hybrid clustering model, the study comparatively analyzed clustering indicators, as shown in Fig. 6. From Fig. 6(a), when the K was 8, the SSE dropped to 5.7×10^6, and the SI reached a peak value of 0.64, indicating that when the K was equal to 8, it was the optimal K value. From Fig. 6(b), the lower the Davies-Bouldin Index (DBI), the better the separation of the clusters (Jayasri, 2022). The Calinski-Harabasz Index (CHI) measures the compactness of samples within clusters and the separation of samples between clusters. Higher values indicate better clustering results (Goswami, 2025). The lowest DBI of the hybrid model was only 0.68, while the DBI of K-means, K-means++ and GMM was 1.25, 1.05 and 0.92, respectively. The hybrid clustering model was reduced by 45.60%, 35.23% and 26.08%, respectively. In addition, the CHI of the hybrid model was 5,861, which was 92.09%, 64.19% and 37.56% higher than that of other clustering methods (3,051, 3,569 and 4,263), respectively. In summary, the hybrid clustering model based on K-means++ and GMM performs well in processing complex data distribution and has better clustering performance.

4.3. Predictive Performance Verification

To verify the prediction performance of the MSTSP user travel prediction model, the study compared it with other benchmark models. Comparative models include MLP, LSTM and Transformer, etc. Fig. 7 presents the prediction performance. In Fig. 7(a), the prediction accuracy of the MSTSP was 94.83%, which was 11.40%, 16.87% and 14.23% higher than the 83.43%, 77.96% and 80.60 of the MLP, LSTM and Transformer models, respectively. In Fig. 7(b), the prediction recall of the MSTSP was 93.87%, which was 9.23%, 10.92% and 9.81% higher than the 85.64%, 82.95% and 84.06% of other models, respectively. From Fig. 7(c), the F1 value of the MSTSP was 94.25%, which was 11.62%, 14.80% and 12.67% higher than the 82.63%, 79.45% and 81.58% of other models, respectively. In summary, the MSTSP model has higher prediction performance and reliability in user travel prediction tasks.

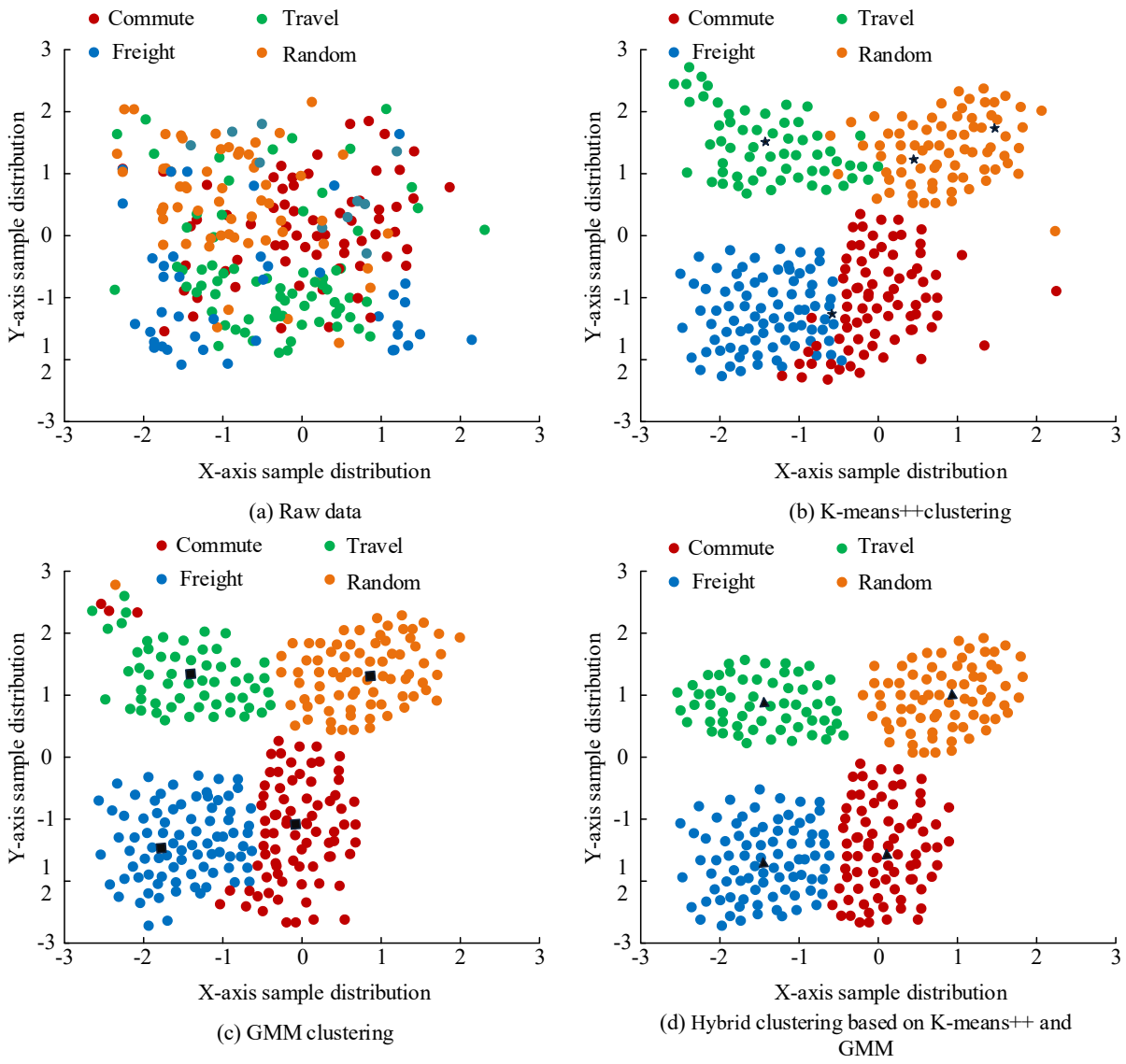


Fig. 5. Visual analysis results

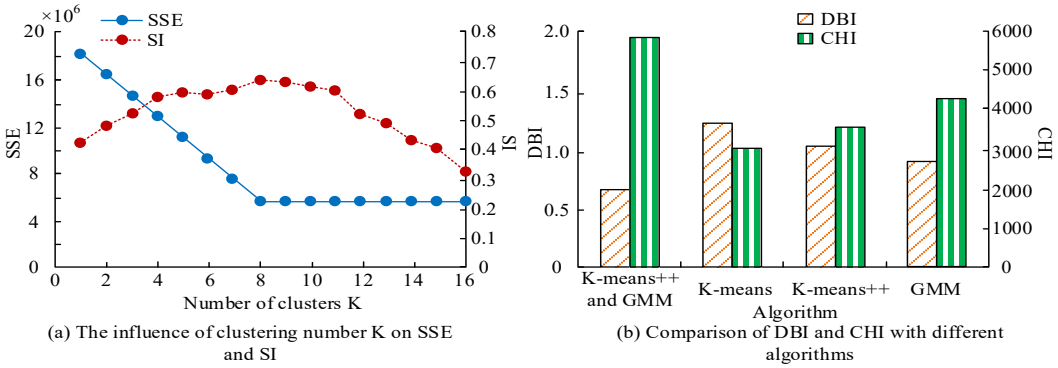
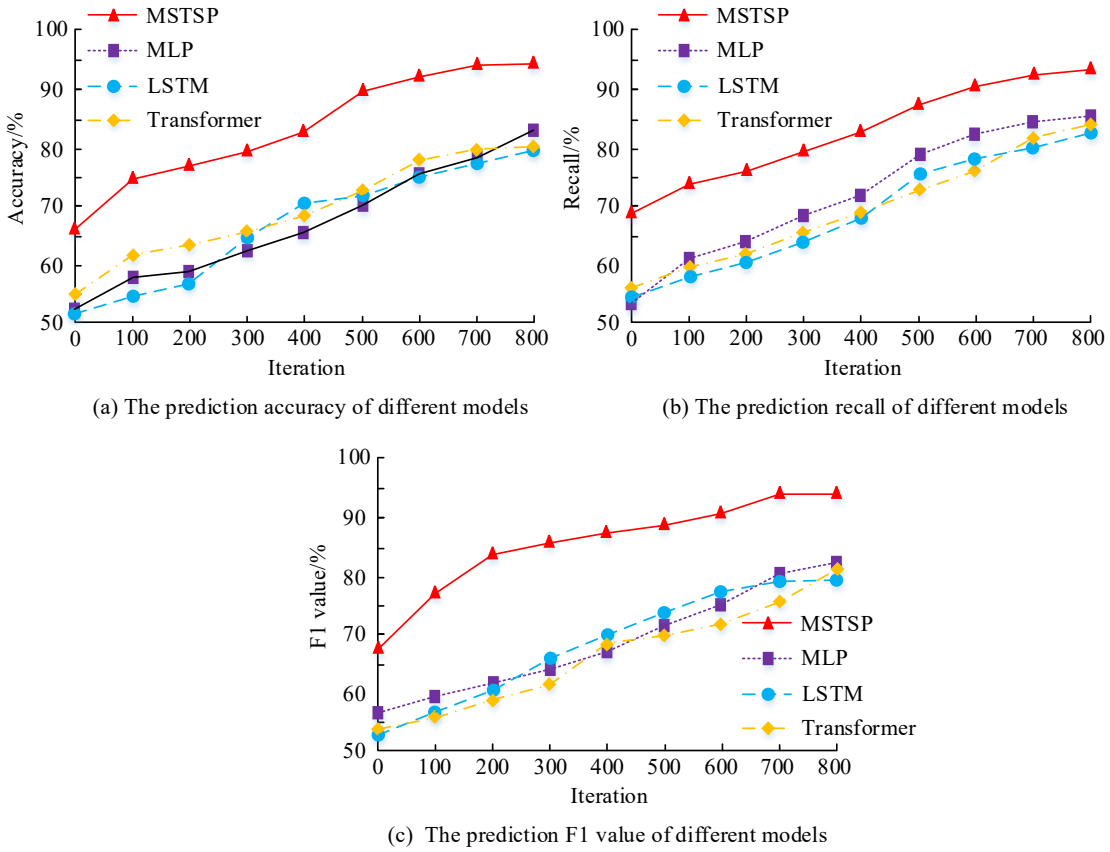


Fig. 6. Comparative analysis of clustering indicators

To further verify the effect of the prediction model, the study added Gaussian noise as random interference and compared the anti-noise interference capability of different prediction models, as shown in Fig. 8. From Fig. 8(a), as the noise intensity increased, the Mean Absolute Error (MAE) value increased. When the noise intensity was 0.2 and 1.2, the MAE values of the MSTSP model were 9.05 and 18.60, respectively, with an increase of 105.42%. In comparison, the MAE of the MLP increased from 15.10 to 31.00, with an increase of 105.30%. The MAE of the LSTM increased from 12.15 to 27.94, with an increase of 129.98%. The MAE of the Transformer increased from 14.08 to 30.25, with an increase of 114.87%. From Fig. 8(b), the Root Mean Square Error (RMSE) of the MSTSP model increased from 13.30 to 27.35, with an increase of 105.64%.

The RMSE of the MLP model increased from 20.25 to 42.05, with an increase of 107.65%. The LSTM model increased from 17.11 to 36.72, with an increase of 114.70%. The Transformer model increased from 18.23 to 38.60, with an increase of 111.63%. In summary, although all models showed an increase in error when adding noise interference, the MSTSP model has the lowest error value and increase, indicating that it has better robustness.

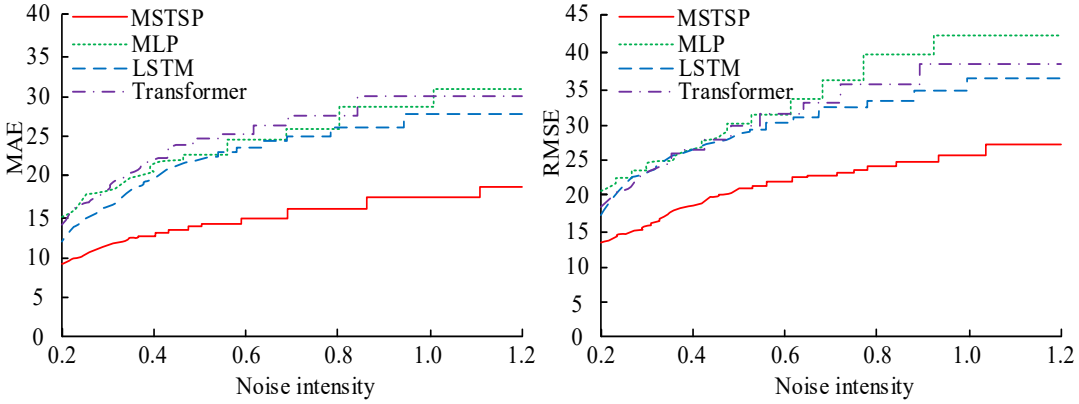


(a) The prediction accuracy of different models

(b) The prediction recall of different models

(c) The prediction F1 value of different models

Fig. 7. Comparison of predictive performance of different models



(a) MAE variation curve with noise intensity

(b) RMSE variation curve with noise intensity

Fig. 8. Comparison of anti-noise interference capability of different prediction models

To verify the role of different modules in the prediction model, each module was added to the network structure to conduct ablation experiments, as presented in Table 3. The accuracy of the full model was 92.15%, and the MAE and RMSE were 8.23 and 12.17, respectively. When the spatial dynamic attention module was removed, the accuracy dropped to 85.32%, the MAE increased to 13.72, and the RMSE increased to 18.93. After removing the temporal attention module, the accuracy was 88.61%, the MAE was 10.50, the RMSE was 15.34. When the cluster feature input was removed, the accuracy dropped significantly to 79.28%, the MAE increased to 15.83, and the RMSE increased to 21.47. In summary, the combined effect of all modules can optimize the prediction performance.

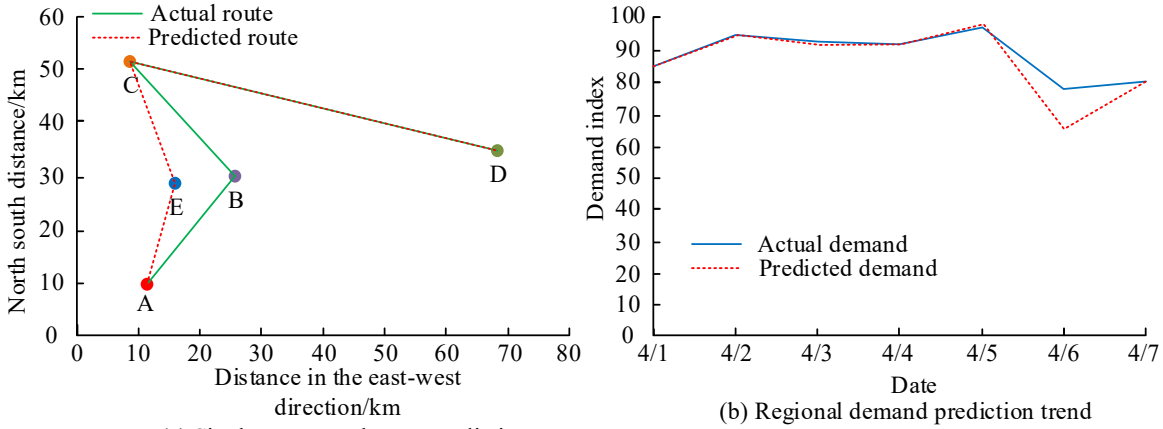
Table 3. Ablation experiment

Model variant	Spatial dynamic attention	Temporal attention	Cluster feature input	Accuracy/%	MAE	RMSE
Full model	√	√	√	92.15	8.23	12.17
Removed spatial Dynamic attention	/	√	√	85.32	13.72	18.93
Removed temporal attention	√	/	√	88.61	10.50	15.34
Removed cluster feature input	√	√	/	79.28	15.83	21.47

4.4. Practical Application Verification

To verify the effect in practical applications, experiments on single-user travel route prediction and regional demand prediction trends were conducted, as presented in Fig. 9. From Fig. 9(a), the MSTSP successfully predicted 75% of the key toll stations in the user's travel path, and only incorrectly predicted station B as station E, resulting in a spatial offset of approximately 16.2km in the path. This error is mainly due to the model's predicted path detouring around Station E in the industrial zone, thereby increasing travel time. From Fig. 9(b), in the seven-day toll station traffic prediction, the model's prediction curve closely matched the actual traffic demand. However, a large error occurred on April 6. At this time, the actual demand and predicted demand were 78 and 66, respectively, with an error of 15.38%. This is because sudden rainfall on that day causes a surge in real demand. In summary, the MSTSP model has shown high effectiveness in practical applications, but it needs further integration of real-time traffic event data to improve robustness in scenarios such as industrial area detours and sudden weather changes.

The actual application efficiency is presented in Table 4. The prediction latency of the MSTSP model was only 45.53ms, significantly lower than that of other prediction models. Its memory usage was only 1.2GB, which was 57.1% and 65.7% lower than that of LSTM and Transformer, respectively. In terms of concurrent processing performance, the MSTSP supported 950 concurrent requests per second, which was 2.5 times that of MLP and 1.5 times that of LSTM, meeting the needs of high-load ETC systems. The data processing time for the MSTSP was only 80.55s, whereas the other models took more than 95s. Although the CPU usage was 65.35%, which was slightly higher than that of the MLP (40.70%), it achieved a balance between computing resources and service performance by achieving the highest throughput and optimal concurrency capabilities. Taken together, the MSTSP model has better efficiency in practical applications.



(a) Single user travel route prediction

(b) Regional demand prediction trend

Fig. 9. Actual application prediction results

Table 4. Comparison of practical application efficiency of different prediction models

Index	MSTSP	MLP	LSTM	Transformer
Prediction latency/ms	45.53	120.42	85.13	110.27
Memory usage/GB	1.2	0.8	2.8	3.5
Supported concurrent users (users/sec)	950	380	620	780
Data processing time/s	80.55	95.16	130.64	115.17
CPU usage/%	65.35	40.70	75.25	70.56
Throughput (requests/sec)	1,205	656	850	1,043

5. Conclusion

This study proposed a hybrid clustering user travel prediction method. The hybrid model performed better when dealing with complex data distributions, with clear, non-overlapping cluster boundaries and effective avoidance of outlier interference. Clustering index analysis showed that the SSE reached its optimum when K was 8, with SI and CHI at 0.64 and 5,861, respectively, and DBI at 0.68, indicating that the clustering effect was significantly better than that of other single methods. The MSTSP achieved 94.83% prediction accuracy, with recall and F1 at 93.87% and 94.25%, respectively, both higher than those of other models. In addition, in the noise interference test, the MSTSP model exhibited the smallest increases in MAE and RMSE, demonstrating strong robustness. Ablation experiments showed that removing spatial dynamic attention, the temporal attention module, or the clustering feature input significantly reduced model performance, with the clustering feature input having the greatest impact on accuracy. In actual application verification, the model successfully predicted 75% of key toll stations, and the 7-day traffic prediction curve was highly consistent with actual demand. In terms of application efficiency, the MSTSP predicted a latency of only 45.53ms and a memory footprint of 1.2GB. It supported 950 users/second concurrent requests, the data processing time was 80.55s, and the throughput was 1,205 requests/second, with outstanding overall performance. In summary, the hybrid clustering ETC big data mining user travel prediction method effectively improves the prediction accuracy and efficiency. However, the research still has the following limitations: Firstly, the data source covers only ETC transaction data from a specific province in China, and road network characteristics and user travel habits are regionally specific. The generalization ability in cross-provincial scenarios needs to be verified; Secondly, the model does not deeply integrate real-time traffic incident data, such as traffic accidents, road construction, temporary control, etc., so there is still room for improvement in the prediction accuracy when dealing with such sudden road conditions; Thirdly, the model retraining needs to be carried out based on the full amount of historical ETC data, and the calculation cost is high, which makes it difficult to quickly adapt to the iterative requirements of dynamic scenarios such as road network structure adjustment and user travel mode change. Therefore, future research can further integrate real-time traffic event data with meteorological data to optimize the adaptability to dynamic scenarios and external factors. Simultaneously, lightweight training methods such as incremental learning and federated learning can be explored to reduce the computational cost of heavy training. In addition, heterogeneous ETC data from multiple provinces can be included for cross-regional validation to enhance the generalization ability and expand its application scope.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The author used ChatGPT solely for language editing and to improve readability. The author reviewed and verified all content and takes full responsibility for the accuracy and integrity of the manuscript.

References

Chen, A. H. L., Liang, Y. C., Chang, W. J., Siauw, H. Y., and Minanda, V. (2022). RFM Model and K-Means Clustering Analysis of Transit Traveller Profiles: A Case Study. *Journal of Advanced Transportation,* 2022(1), 1108105-1108109. doi: 10.1155/2022/1108105

Chen, X., Cheng, Z., Jin, J. G., Trépanier, M., and Sun, L. (2023). Probabilistic forecasting of bus travel time with a Bayesian Gaussian mixture model. *Transportation Science,* 57(6), 1516-1535. doi: 10.1287/trsc.2022.0214

Chango, W., Lara, J. A., Cerezo, R., and Romero, C. (2022). A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(4), 1458-1463. doi: 10.1002/widm.1458

Dou, H., Liu, Y., Chen, S., Zhao, H., and Bilal, H. (2023). A hybrid CEEMD-GMM scheme for enhancing the detection of traffic flow on highways. *Soft Computing,* 27(21), 16373-16388. doi: 10.1007/s00500-023-09164-y

Gemma, A., Mannini, L., Crisalli, U., and Cipriani, E. (2024). Improving Urban Travel Time Estimation Using Gaussian Mixture Models. *IEEE Transactions on Intelligent Transportation Systems,* 25(10), 14154-14163. doi: 10.1109/tits.2024.3390792

Gopinathan, K. N., Murugesan, P., and Jeyaraj, J. J. (2024). Stock price prediction using a novel approach in Gaussian mixture model-hidden Markov model. *International Journal of Intelligent Computing and Cybernetics,* 17(1), 61-100. doi: 10.1108/IJICC-03-2023-0050

Goswami, S. S., Mondal, S., Sarkar, S., Gupta, K. K., Sahoo, S. K., and Halder, R. (2025). Artificial Intelligence-Enabled Supply Chain Management: Unlocking New Opportunities and Challenges. *Artificial Intelligence and Applications*, 3(1), 110-121. doi: 10.47852/bonviewAIA42021814

Jayasri, N. P., and Aruna, R. (2022). Big data analytics in health care by data mining and classification techniques. *ICT Express,* 8(2), 250-257. doi: 10.1016/j.icte.2021.07.001

Lai, X., Zhang, S., Mao, N., Liu, J., and Chen, Q. (2022). Kansei engineering for new energy vehicle exterior design: An internet big data mining approach. *Computers & Industrial Engineering,* 165(1), 107913-107920. doi: 10.1016/j.cie.2021.107913

Lakshmi, P. S., Saxena, M., Koli, S., Joshi, K., Abdullah, K. H., and Gangodkar, D. (2022). Traffic response system based on data mining and internet of things (IoT) for preventing accidents. *2022 2nd International Conference On Advance Computing And Innovative Technologies In Engineering,* 16(1), 1092-1096. doi: 10.1109/iCacite53722.2022.9823923

Oatley, G. C. (2022). Themes in data mining, big data, and crime analytics. Wiley Interdisciplinary Reviews: *Data Mining*

and Knowledge Discovery, 12(2), 1432-1437. doi: 10.1002/widm.1432

Park, H. S., Park, Y. W., Kwon, O. H., and Park, S. H. (2022). Applying Clustered KNN Algorithm for Short-Term Travel Speed Prediction and Reduced Speed Detection on Urban Arterial Road Work Zones. *Journal of Advanced Transportation,* 2022(1), 1107048-1107057. doi: 10.1155/2022/1107048

Sarwar, T., Seifollahi, S., Chan, J., Zhang, X., Aksakalli, V., Hudson, I., and Cavedon, L. (2022). The secondary use of electronic health records for data mining: data characteristics and challenges. *ACM Computing Surveys (CSUR),* 55(2), 1-40. doi: 10.1145/3490234

Saygılı, A. (2022). Computer-aided detection of COVID-19 from CT images based on Gaussian mixture model and kernel support vector machines classifier. *Arabian Journal for Science and Engineering,* 47(2), 2435-2453. doi: 10.1007/s13369-021-06240-z

Shahrier, M., Hasnat, A., Al-Mahmud, J., Huq, A. S., Ahmed, S., and Haque, M. K. (2024). Towards intelligent transportation system: A comprehensive review of electronic toll collection systems. *IET Intelligent Transport Systems,* 18(6), 965-983. doi: 10.1049/itr2.12500

Yang, Y., Yuan, Z., and Meng, R. (2022). Exploring traffic crash occurrence mechanism toward cross-area freeways via an improved data mining approach. *Journal of transportation engineering, Part A: Systems*, 148(9), 4022052-4022059. doi: 10.1061/jtepbs.0000698

Yu, B., Mao, W., Lv, Y., Zhang, C., and Xie, Y. (2022). A survey on federated learning in data mining. Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery*, 12(1), 1443-1448. doi: 10.1002/widm.1443

Xie, L., Guo, T., Chang, J., Wan, C., Hu, X., Yang, Y., and Ou, C. (2023). A novel model for ship trajectory anomaly detection based on Gaussian mixture variational autoencoder. *IEEE Transactions on Vehicular Technology,* 72(11), 13826-13835. doi: 10.1109/TVT.2023.3284908

Zhang, P., Zheng, J., Lin, H., Liu, C., Zhao, Z., and Li, C. (2023). Vehicle trajectory data mining for artificial intelligence and real-time traffic information extraction. *IEEE Transactions on Intelligent Transportation Systems,* 24(11), 13088-13098. doi: 10.1109/tits.2022.3178182

Zhang, Y., and Sun, Y. (2024). Smart travel planning system based on kernel density estimation and similarity metric clustering algorithm. *Evolutionary Intelligence*, 17(5), 4227-4238. doi: 10.1007/s12065-024-00980-1

Dandan Wang received her bachelor's degree in communication engineering from Shandong University of Science and Technology in 2014 and her master's degree from Shanghai Maritime University in 2016. Her teaching career includes positions at Yantai Nanshan University (2017–2019), Zibo Vocational Institute (2019–2025), and Zibo Polytechnic University, where she has been a faculty member since 2025. She has authored one monograph, holds two software copyrights, has led five municipal-level or higher research projects, and has participated in nine provincial-level or higher projects.