Applied Artificial Intelligence Research

INNOVIAIR

A Case Study of the Influence of Multifarious Factors on Traffic Flow Forecasting

Xiaojun Wang[a,*]

[a]College of Computer Science and Technology, Qingdao University, Qingdao, Shandong 266071, China

ARTICLEINFO

Keywords:
Traffic Flow Forecasting
Multi-Source Data Fusion
Principal Component Analysis
Random Forest
SHAPley Additive ExPlanations

ABSTRACT

Accurate traffic flow forecasting serves as the core foundation for intelligent transport systems to achieve efficient traffic management and optimized resource allocation, holding significant importance for alleviating congestion in modern cities. Influenced by the complex interplay of diverse heterogeneous factors such as meteorological conditions, temporal cycles, and unforeseen events, urban traffic flow data exhibits pronounced nonlinearity and random fluctuations, rendering high-precision forecasting exceptionally challenging. While mainstream deep learning models have advanced prediction accuracy, they struggle to quantify the specific contributions of different factors and often overlook significant multicollinearity issues among multi-source data. Addressing these challenges, this paper introduces several improvements: Firstly, it constructs a fused dataset incorporating multidimensional external factors such as meteorological conditions and events; secondly, it proposes an explainable prediction framework based on Random Forest with raw feature analysis by Principal Component Analysis (PCA); Thirdly, the SHAP game-theoretic method is introduced to achieve transparent attribution of prediction results. This paper first employs PCA to extract principal components from high-dimensional multi-source factors, effectively eliminating multicollinearity in the data. Subsequently, a robust random forest regression model is constructed for prediction. Based on four independent datasets from Beijing's TaxiBJ service, the proposed framework undergoes comprehensive performance validation and analysis. Results demonstrate that the model maintains a coefficient of determination consistently above 0.85 across all annual datasets, exhibiting outstanding predictive accuracy and robustness across temporal cycles. SHAP analysis further reveals a stable decision mechanism characterized by the first principal component driving periodicity, with subsequent components providing dynamic fine-tuning, successfully achieving traffic flow prediction that combines high precision with strong interpretability.

1.Introduction

1.1.Traffic flow forecasting

With the acceleration of urbanization, traffic congestion has become one of the most pressing challenges facing modern cities. Intelligent Transportation Systems serve as a key solution to this problem, with one of their core functions being the precise forecasting of traffic flow[1]. Traffic flow forecasting aims to predict future traffic conditions during specific time periods based on historical observation data. Accurate forecasting not only assists managers in implementing effective traffic management and signal control but also provides travellers with optimal route planning. However, traffic flow data exhibits highly random and non-linear characteristics, influenced by the interplay of numerous complex factors, rendering high-precision forecasting exceptionally challenging. While existing deep learning models have achieved significant advances in predictive accuracy, they often lack interpretability regarding forecast outcomes, making it difficult to quantify the specific contributions of different factors[2].

1.2.Multi-source factors influencing traffic flow forecasting

Traffic flow does not exist in isolation; it is the result of multiple factors acting in concert. In this study, we categorize

* Corresponding author.
E-mail address: wangxiaojun@ubinet.cn.
https://doi.org/10.65455/sbd3t486
Received 12 May 2026; Received in revised form 28 May 2026; Accepted 3 June 2026; Available online 10 June 2026

the factors influencing traffic flow into the following groups. These factors often exhibit highly nonlinear interactions, and they have been demonstrated to constitute key components of complex urban systems[3].

Weather conditions directly affect road capacity and driver behaviour, leading to reduced speeds or abnormal traffic flow[4]. Traffic flow exhibits significant periodicity, influenced by temporal factors such as commuting peaks and differences between weekdays and weekends. Sudden incidents, public holidays or large-scale gatherings can disrupt normal traffic patterns, triggering non-periodic congestion. The distribution of points of interest in the surrounding area and the topological structure of the road network determine the spatial dependence of traffic flow. Traffic volume, speed and occupancy rates over a recent period provide the most direct basis for forecasting future conditions.

1.3.Limitations of existing research and issues to be addressed

With the advancement of artificial intelligence technology, traffic flow forecasting research has evolved from early parametric statistical models such as ARIMA[5] and Kalman filtering[6] to the currently dominant deep learning models. Most existing research focuses on leveraging CNNs[7], LSTM networks and their variants[8], or more advanced GCNs[9] to deeply mine the spatiotemporal dependencies within traffic data. By stacking complex network layers, these methods have indeed achieved significant breakthroughs in prediction accuracy, enabling them to capture the nonlinear characteristics of traffic flow effectively.

However, despite the existing models demonstrating excellent performance in terms of predictive accuracy, two significant limitations remain in practical applications: To improve accuracy, existing research tends to continuously increase the dimensionality of input features, resulting in underutilisation of data. However, there is often a high degree of correlation between data from multiple sources. Feeding this high-dimensional, redundant data directly into models not only increases the computational burden but may also interfere with the model's identification of key drivers due to multicollinearity[10]. Real-world traffic data frequently contains sensor errors or sudden outliers, yet many models lack sensitivity to data quality. The internal structures of most deep learning models are extremely complex, lacking transparent decision-making mechanisms. Although models can provide accurate predictive values, they are unable to explain the reasons behind the predictions or identify the specific factors leading to congestion. This lack of interpretability makes it difficult for traffic managers to formulate targeted congestion mitigation strategies based on model results[11].

The core issue this study aims to address is: How to clearly separate and quantify the specific contributions of different source factors to traffic flow forecasting whilst ensuring high predictive accuracy.

1.4.Contribution

To address the aforementioned issues, this study proposes an analytical framework integrating machine learning benchmark models with multi-source data fusion. The principal contributions are as follows:

1.Analysing multi-source heterogeneous datasets: This study integrates traffic flow data, meteorological data, date data and event data, supplemented by local real-world data for validation, comprehensively covering the key external characteristics influencing traffic flow.

2.Establish interpretable strong benchmarks: Random forest regression was introduced as the core baseline model. Leveraging its ability to capture non-linear relationships, robustness to outliers, and the advantage of requiring no complex preprocessing, a competitive predictive baseline was established.

3.Quantifying the Importance of Influencing Factors: By leveraging the built-in feature importance assessment capability of random forests, coupled with subsequent SHAP analysis methods, factors such as historical traffic volumes, weather conditions, and POI information were quantitatively ranked and interpreted. This approach not only achieved high-precision forecasting but also provided an explainable basis for understanding the operational mechanisms of urban traffic systems.

2.Methodology

2.1.Random forest regression

In this study, we selected random forest regression as the benchmark model and core interpretability analysis tool. Random forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to produce the final forecast. Its core principle lies in combining multiple decision trees to form a robust model, thereby significantly enhancing overall generalization capability and robustness[12].

2.1.1.Model principles and construction mechanisms

The construction of a Random Forest relies primarily on two key stochastic processes: Bootstrap Aggregating and Random Feature Selection. This dual stochastic mechanism effectively reduces the risk of model overfitting and renders the model insensitive to noise in the data. By using bootstrap sampling with replacement, multiple distinct subsets are generated from the original training dataset[13], with each subset used to train an independent decision tree; when splitting at each node of the decision tree, the model does not select the optimal split point from all features, but rather from a randomly selected subset of features[14].

2.1.2.Adaptive analysis

Random Forest regression models demonstrate a high degree of adaptability to the complex task of traffic flow prediction. Traffic flow is influenced by a combination of multiple factors, and these relationships are often highly non-linear[15]. As an ensemble method based on decision trees,

Random Forest is naturally adept at capturing such complex non-linear interactions and handling intricate non-linear relationships. Due to the use of ensemble and averaging strategies, the model exhibits good robustness against outliers in the training data, which is crucial for handling real-world traffic data containing sensor noise or sudden disturbances. Unlike data-scale-sensitive models such as Support Vector Machines[16], Random Forests do not require strict data standardisation or normalisation, thereby simplifying the pre-processing workflow.

2.1.3.Model character

In this study, the Random Forest serves as an analytical tool. Using its built-in feature importance assessment function, the model calculates the average reduction in impurity achieved when each feature is used for splitting across all trees. We will use this to quantify the contribution of factors such as historical traffic volume, weather conditions and POI information to predictive accuracy, and to rank the factors influencing traffic flow, thereby clearly distinguishing the specific contributions of factors from different sources.

2.2.The methodology and technical approach of this paper

In this research, PCA is firstly applied to test multicollinearity and analyze principal component loadings of high-dimensional multi-source heterogeneous features, revealing feature redundancy and core driving factors. Original multi-source features are used directly as model inputs with no dimensionality reduction. A robust random forest regression model is established for prediction. PCA only functions for feature interpretation and mechanism analysis, rather than dimensionality transformation of model inputs. Similarly, in traffic-related forecasting tasks, PCA has proven to be an efficient method for handling multicollinearity in multi-source data, and when combined with sequence models, it can further enhance predictive performance[17].

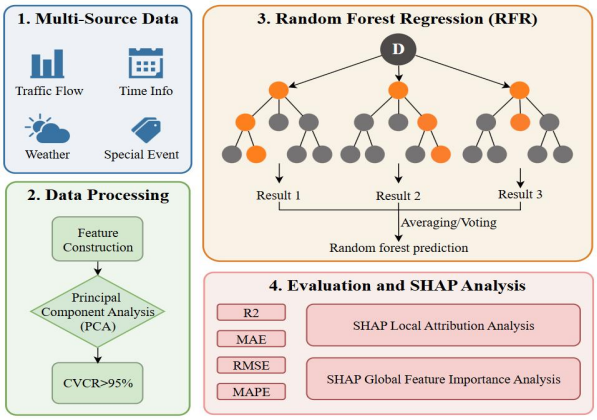


Fig 1. The flow diagram of this research

3.Experiment

3.1.Dataset and experimental environment

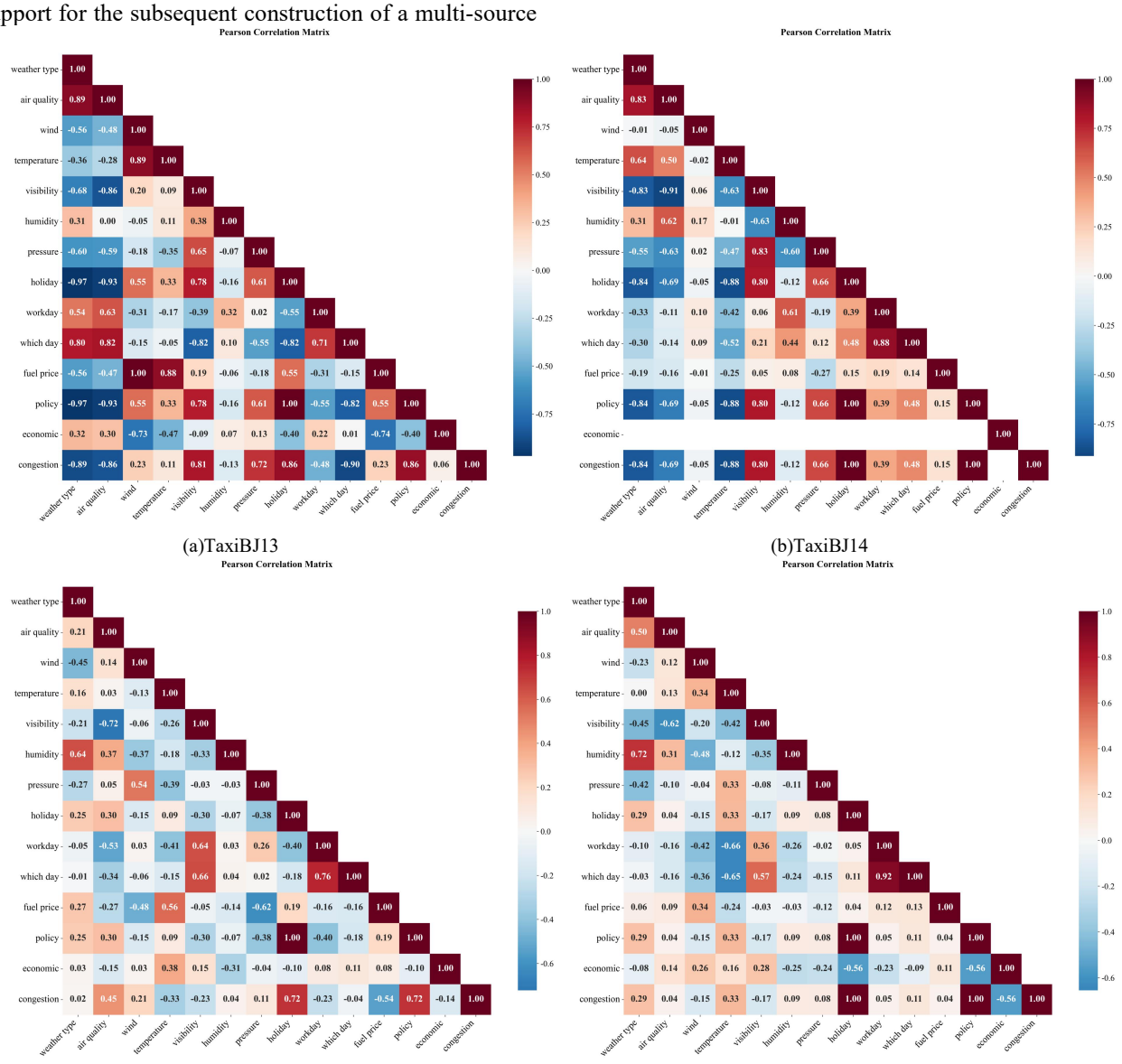
To comprehensively evaluate the impact of multiple factors on traffic flow forecasting, our experiment utilized the TaxiBJ dataset comprising Beijing taxi flow data from 2013 to 2016, provided by Zhang et al.[18]. The urban boundaries of Beijing have expanded year by year, with the functional positioning of different districts within the city evolving accordingly. It is therefore necessary to conduct an annual analysis of taxi traffic volumes across each year in Beijing. Organized by year, these four datasets are named TaxiBJ13, TaxiBJ14, TaxiBJ15 and TaxiBJ16, respectively.

We have divided the central urban area of Beijing into a 32×32 grid, with each grid cell representing a distinct area. The volume of taxis leaving each area serves as an indicator of urban travel demand within that area. The sampling interval for taxi traffic is set at 30 minutes, resulting in a total of 48 time slots per day. Maximum/minimum demand refers to the highest and lowest levels of travel demand recorded within the respective scopes of the four datasets: TaxiBJ-13, TaxiBJ-14, TaxiBJ-15 and TaxiBJ-16. These four datasets reflect seasonal variations. Furthermore, the natural environmental factors included in the datasets comprise weather type, wind speed, air temperature and atmospheric pressure; socio-economic factors include public holidays/non-holidays, working days/rest days, different days of the week, fuel prices, transport policies, traffic congestion levels, the pandemic, residents' income levels and the city's overall economic situation.

Within the time frame covered by the TaxiBJ dataset, for factors such as fuel prices, transport policies, the pandemic, household income levels and the city's overall economic situation, we have used the 'abnormal dates' corresponding to these influencing factors to replace their actual values or categories. For fuel prices, if the price on a given day deviates from the national average natural gas price by more than two standard deviations, that day is classified as an abnormal fuel price day; for transport policies, the dates on which the Beijing Municipal Commission of Transport implemented city-wide taxi restrictions in the central urban area are classified as abnormal transport policy days; for the pandemic, dates on which city-wide taxi restrictions in the central urban area were imposed due to infectious disease outbreaks are classified as abnormal pandemic days; For household income levels and the city's overall economic situation, if the total household income or per capita income in Beijing on a given day deviates from the corresponding average by more than two standard deviations, that day is classified as an 'abnormal day' for household income or the city's economic situation. To protect data privacy, the traffic congestion indicator does not utilise actual traffic congestion data, but instead employs anonymised traffic density index values provided by the DataFountain platform[19]. Traffic density refers to the number of vehicles per lane, specifically the number of vehicles per unit length of lane at a given moment; the traffic density index is calculated by multiplying traffic density by a positive real number. Due to data privacy policies, the DataFountain platform does not disclose the specific value of this positive real number. All data is closely linked to individual grids.

To preliminarily explore the interrelationships among multi-source data, we calculated the Pearson correlation coefficients for each primary feature in the TaxiBJ dataset[20], with results presented in Figure 2. The figure demonstrates that congestion exhibits significant correlations with factors

such as weather type and holiday status, thereby providing data support for the subsequent construction of a multi-source fusion prediction model.



(a)TaxiBJ13 (b)TaxiBJ14

(c)TaxiBJ15 (d)TaxiBJ16

Fig 2. Pearson correlation coefficient matrix for multi-source features in the TaxiBJ dataset

3.2.Data design

This study defines the traffic flow forecasting task as follows: utilizing historical observational data spanning the past T time slots, combined with known external factors at the same future time point, to predict traffic flow for the future τ time slot.

To make full use of multi-source information, the characteristics of each spatial node—such as traffic, speed and occupancy—are organised into a three-dimensional tensor of the form [N, T, F], where N denotes the number of nodes, T denotes the time step, and F denotes the number of features. To capture spatial dependencies, an adjacency matrix is constructed based on the actual road network distances between nodes. This is then normalized using a thresholded Gaussian kernel to define the spatial graph structure. External factors corresponding to each future prediction time point shall be encoded and concatenated into a feature vector. This vector shall undergo deep integration with the spatio-temporal sequence at the feature dimension level, thereby assisting the model in capturing non-periodic traffic fluctuations.

To evaluate the extent to which principal components retain original information and determine the optimal number of principal components, we employed Contribution Rate and Cumulative Variance CR as assessment metrics[21], as illustrated in Figure 3.

$$CVCR=\sum_{i=1}^{n} CR_i \quad (1)$$

Here, CR denotes the proportion of variance attributable to a single common factor relative to the total variance, serving to assess the magnitude of that factor's influence on the dependent variable. Conversely, CVCR represents the proportion of variance explained by the sum of all extracted common factors relative to the total variance, reflecting the collective explanatory power of all factors on the dependent variable. CVCR may be expressed as the sum of all selected principal component CR values, calculated as shown in Equation (1). Regarding evaluation criteria, CVCR values greater than 0.5 are generally considered acceptable, values greater than or equal to 0.7 are deemed satisfactory, values

greater than or equal to 0.8 are regarded as ideal, and values greater than or equal to 0.95 are regarded as excellent[22].

As indicated by the dashed line in Figure 3, we set the extremely optimal threshold line at 0.95. Experimental results indicate that on the BJ13 and BJ14 datasets, the CVCR values of the first five principal components rapidly increased and all exceeded the 0.95 threshold, with some approaching 100%. On the BJ15 and BJ16 datasets, the CVCR values of the first seven principal components satisfied the preset 0.95 threshold. This implies that retaining only the first three principal components suffices to explain nearly all variance in the original data, reflecting the vast majority of information contained within the original variables. In other words, the projection of the original variables onto the feature vectors PC1 to PC7 can be regarded as an effective new variable capable of replacing the original high-dimensional data, indicating significant linear correlations within the original 14-dimensional feature space. This study always uses raw multi-source heterogeneous features as model inputs and does not employ any PCA dimensionality-reduced variables. The above PCA results only verify the multicollinearity and redundancy of original features, providing a theoretical basis for the subsequent SHAP analysis of original features.

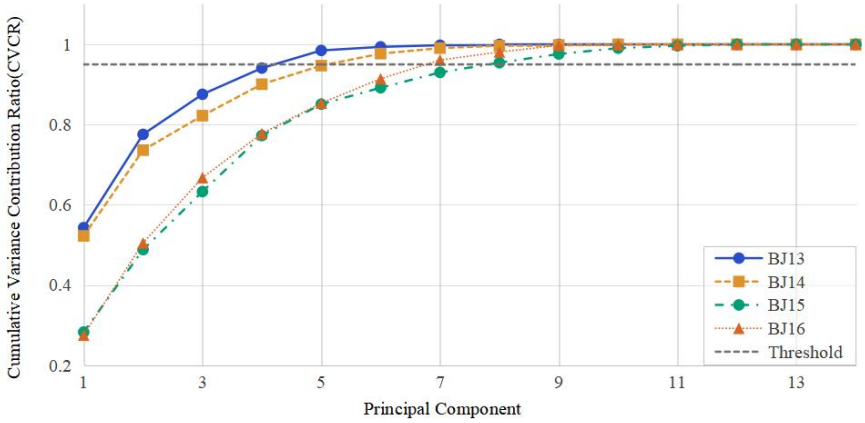


Fig 3. The CVCR curves for datasets BJ13, BJ14, BJ15, and BJ16

Table 1. Top-5 loading features and corresponding transportation domain categories of PC1 across four datasets

Dataset	Principal Component	Original Feature	Loading Coefficient	Feature Category
BJ13	PC1	holiday	0.3582	Temporal Cycle
BJ13	PC1	policy	0.3582	Traffic Incident
BJ13	PC1	weather type	-0.3507	Meteorology
BJ13	PC1	air quality	-0.3470	Meteorology
BJ13	PC1	congestion	0.3172	Traffic State
BJ14	PC1	visibility	0.4055	Meteorology
BJ14	PC1	holiday	0.3760	Temporal Cycle
BJ14	PC1	policy	0.3679	Traffic Incident
BJ14	PC1	weather type	-0.3663	Meteorology
BJ14	PC1	congestion	0.3597	Traffic State
BJ15	PC1	holiday	0.3971	Temporal Cycle
BJ15	PC1	policy	0.3971	Traffic Incident
BJ15	PC1	workday	-0.3871	Temporal Cycle
BJ15	PC1	visibility	-0.3695	Meteorology
BJ15	PC1	air quality	0.3282	Meteorology
BJ16	PC1	holiday	0.4556	Temporal Cycle
BJ16	PC1	policy	0.4556	Traffic Incident
BJ16	PC1	congestion	0.4556	Traffic State
BJ16	PC1	economic	-0.3018	Economic
BJ16	PC1	weather type	0.2725	Meteorology

We conducted PCA component loading analysis on four independent annual datasets (BJ13, BJ14, BJ15, BJ16) separately, and restored each principal component to its corresponding original traffic characteristics.

The loading coefficient reflects the correlation degree between each original feature and the principal component, with a larger absolute value representing a stronger contribution. The Table 1 show that the feature composition of principal components is highly consistent across four years:

The PC1 (top-loading features) are consistently derived from temporal cycle, traffic incident, and meteorology across all four datasets. It explains 12%–13% of the total variance and serves as the core comprehensive driving factor of urban traffic flow, which determines the baseline value of model prediction. PC2–PC7 correspond to economic–meteorological disturbance, meteorology-dominated disturbance, and local fine-tuning factors respectively.

3.3.Baseline and evaluation indicator design

3.3.1.Baseline model

To comprehensively validate the effectiveness and superiority of the proposed PCA-random forest fusion prediction framework, this study introduces several representative traditional and fundamental models from the field of traffic flow time series forecasting as comparative baselines:

Historical Average: A widely employed heuristic benchmark model. It calculates forecast values by determining the average flow rate for the corresponding period in historical records. This model relies primarily on historical periodicity and is unable to capture non-linear fluctuations arising from sudden events or meteorological changes[23].

ARIMA: An early parametric statistical model. While possessing mathematical rigour in handling stationary time series, it struggles to effectively integrate heterogeneous external factors such as meteorological data and events from multiple sources, and exhibits limited capability in fitting highly non-linear traffic flows.

Support Vector Regression: Unlike the random forest employed in this study, SVR exhibits extreme sensitivity to data scaling. When processing high-dimensional multi-source features without rigorous dimensionality reduction and scaling, it is highly prone to local optima[8].

3.3.2.Evaluation indicators

To comprehensively evaluate prediction accuracy from multiple perspectives, this study employs three widely used statistical metrics: mean absolute error, root mean square error, and mean absolute percentage error. All metrics are calculated across all detectors and all time slices within the test set, with lower values indicating greater prediction accuracy.

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (2)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (3)$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \quad (4)$$

3.4.Hyperparameter tuning process

For the Random Forest model, as it involves two core stochastic processes, we employ a grid search strategy combined with cross-validation to determine the optimal combination of parameters. The number of decision trees was varied across the range [10, 50, 100, 200, 500] to balance computational cost with the model's generalization ability. The maximum number of features was controlled by adjusting the size of the feature subsets considered during node splitting. The tree depth was limited to prevent the model from overfitting on the training set.

For deep learning models, we primarily adjusted the learning rate, batch size and the number of layers in the convolutional layers, and employed early stopping to terminate training when the validation set error ceased to decrease, thereby obtaining the optimal model weights. This optimization process ensured that all models included in the comparison were operating at peak performance, thus guaranteeing the fairness and validity of the experimental results.

4.Result

4.1.Quantitative analysis

In order to comprehensively evaluate the predictive performance and generalization ability of the Random Forest regression model across different traffic conditions, we conducted independent tests on four standard datasets covering different time periods.

Table 2. Statistics on the model's predictive performance across different years of the TaxiBJ dataset.

	R^2	MAE	RMSE	MAPE
BJ13	0.8558	12.44	15.88	27.79%
BJ14	0.8800	15.54	18.30	30.90%
BJ15	0.8571	15.96	19.37	28.08%
BJ16	0.8600	13.55	15.95	33.94%

The experimental results indicate that the Random Forest model exhibits good robustness. Firstly, the R^2 of the model remains consistently above 0.85 across all datasets, reaching a maximum of 0.88, demonstrating its ability to effectively capture the primary patterns of variation in traffic flow. Secondly, in terms of error metrics, the BJ13 dataset performed best, achieving the lowest MAE and RMSE. It is worth noting that although the absolute error for BJ16 is small, its MAPE is relatively high; this is primarily due to the large proportion of low-traffic samples in this dataset, which makes the relative error more sensitive to changes in values. Overall, the model's consistent performance across years validates its adaptability under different traffic patterns.

4.2.Qualitative analysis

To assess more intuitively the model's ability to capture the time-varying characteristics of traffic flow—particularly its performance in identifying traffic congestion peaks and

sudden changes—we selected representative consecutive time segments from each of the four test datasets (BJ13, BJ14, BJ15 and BJ16) and plotted comparison curves of the observed traffic flow data against the model's predicted values, as shown in Figures 4(a)–(d).

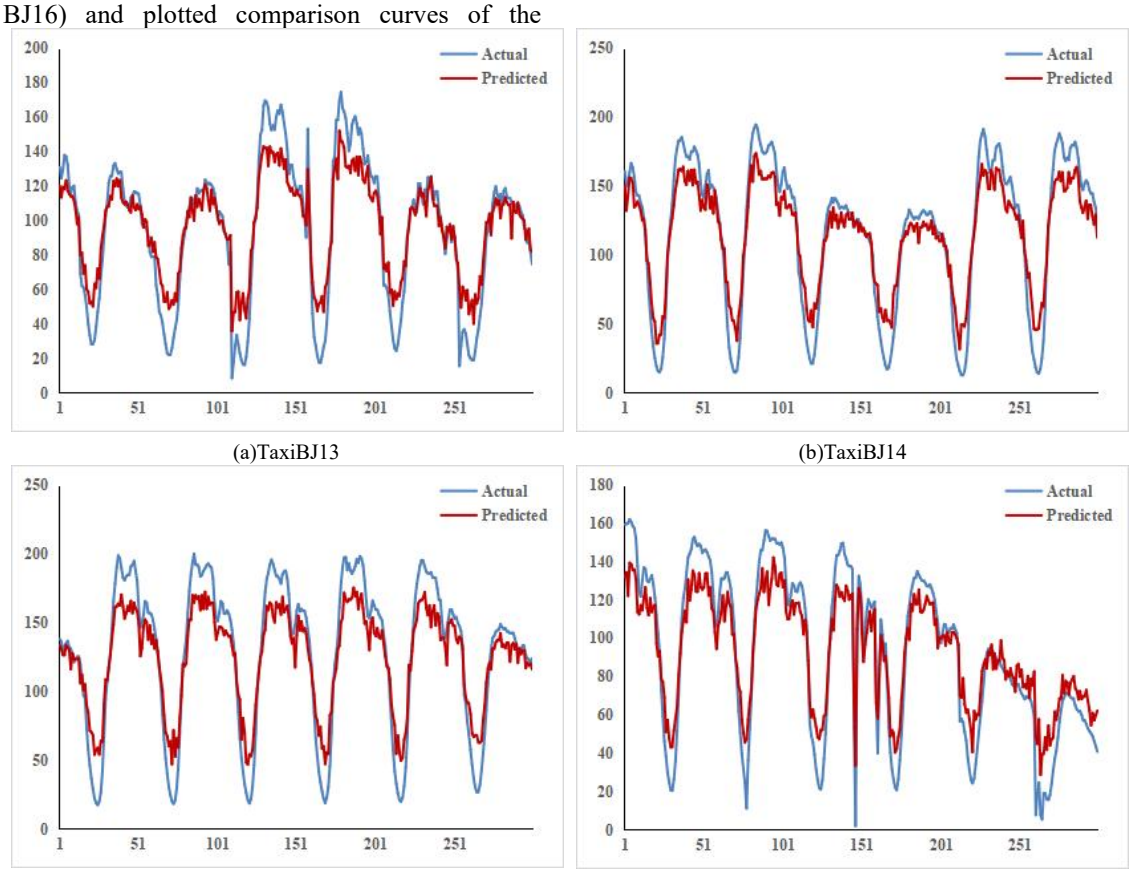


Fig 4. Time-series comparison of ground truth vs. predicted traffic flow on the BJ13-BJ16 datasets

An examination of the forecast curves reveals that, across different datasets spanning four years, the forecast curves closely track the fluctuations of the actual curves. The model has successfully captured the significant periodic characteristics of urban traffic flow, namely the distinct morning and evening peak structures and the night-time trough period. In contrast, whilst traditional statistical models can approximate the general trend, they often exhibit a noticeable phase lag during periods of abrupt traffic changes, such as the transition between weekdays and weekends. This consistent performance across datasets provides intuitive validation of the model's strong generalisation ability.

The core challenge in traffic forecasting lies in accurately predicting the start and end times, as well as the peak intensity, of the morning and evening rush hours. When processing multi-source data that includes meteorological factors or unexpected events, SVR and HA models often exhibit a smoothing effect on extreme values, resulting in severely underestimated forecasts for peak periods. As can be seen in Figure 4, the model exhibits high response sensitivity during both the morning rush hour, when traffic volume rises rapidly, and the evening rush hour, when it declines gradually, with no significant phase lag observed. Although there may be a slight underestimation of predicted values at certain extreme peak moments, overall, the model can accurately define the scope of congestion periods.

The model also provides a good fit for variations in the curve's shape during off-peak hours and across different days. This indicates that the Random Forest regression model, aided by the incorporation of multi-source features, is capable of effectively modelling the complex non-linear relationships inherent in traffic flow, rather than merely performing simple linear extrapolation. Furthermore, the model demonstrates excellent robustness in handling data noise and local disturbances. As can be seen in Figures 4(a)–(d), real-world traffic data often contains a certain degree of random noise. However, the predicted curves do not exhibit severe overfitting-induced fluctuations as a result, but instead maintain a relatively smooth trend-following pattern.

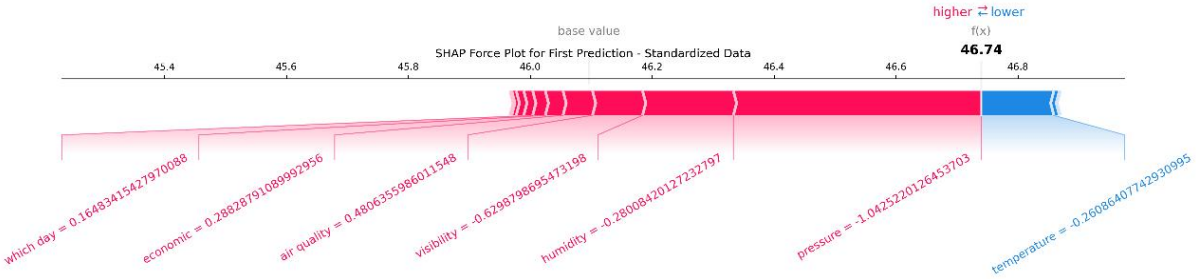
4.3.Analysis of influencing factors

Combined with the PCA component loading results of four datasets in Section 3.2, we correspond the principal component information to the original transportation features. To gain a deeper understanding of how the model utilizes raw input features to make decisions across four different datasets, this study employs the SHAPley Additive exPlanations method[24]. Based on game theory, SHAP provides additive explanations for the predicted values of each sample.

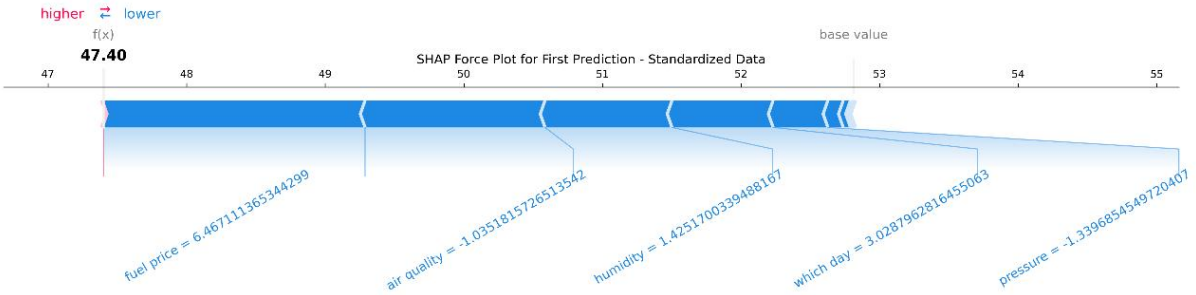
4.3.1.Force plot analysis

We randomly selected representative prediction samples from the test sets of the four datasets and plotted SHAP Force Plots, as shown in Figures 5(a)–(d). In the figures: the Base Value represents the model's average predicted output across the entire dataset; the Output Value represents the model's final prediction for the specific sample in question; the red

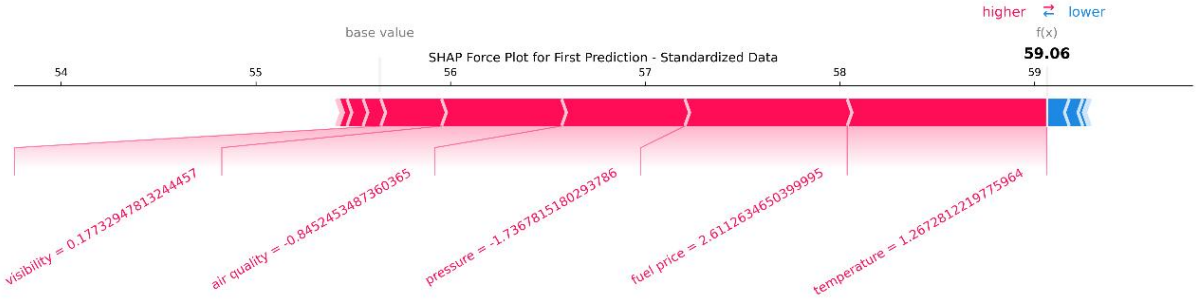
bars indicate that the feature pushes the predicted value upwards; blue bars indicate that the feature pulls the predicted value downwards; the length of the bars intuitively reflects the absolute magnitude of the feature's influence on the prediction result.



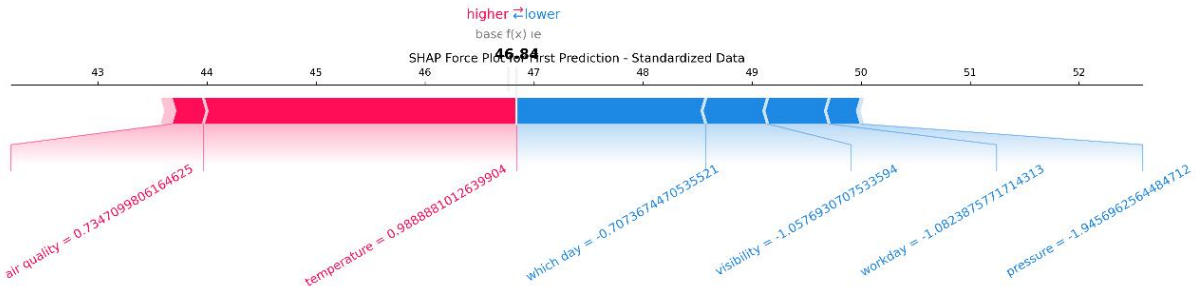
(a)BJ13 SHAP force plot



(b)BJ14 SHAP force plot



(c)BJ15 SHAP force plot



(d)BJ16 SHAP force plot

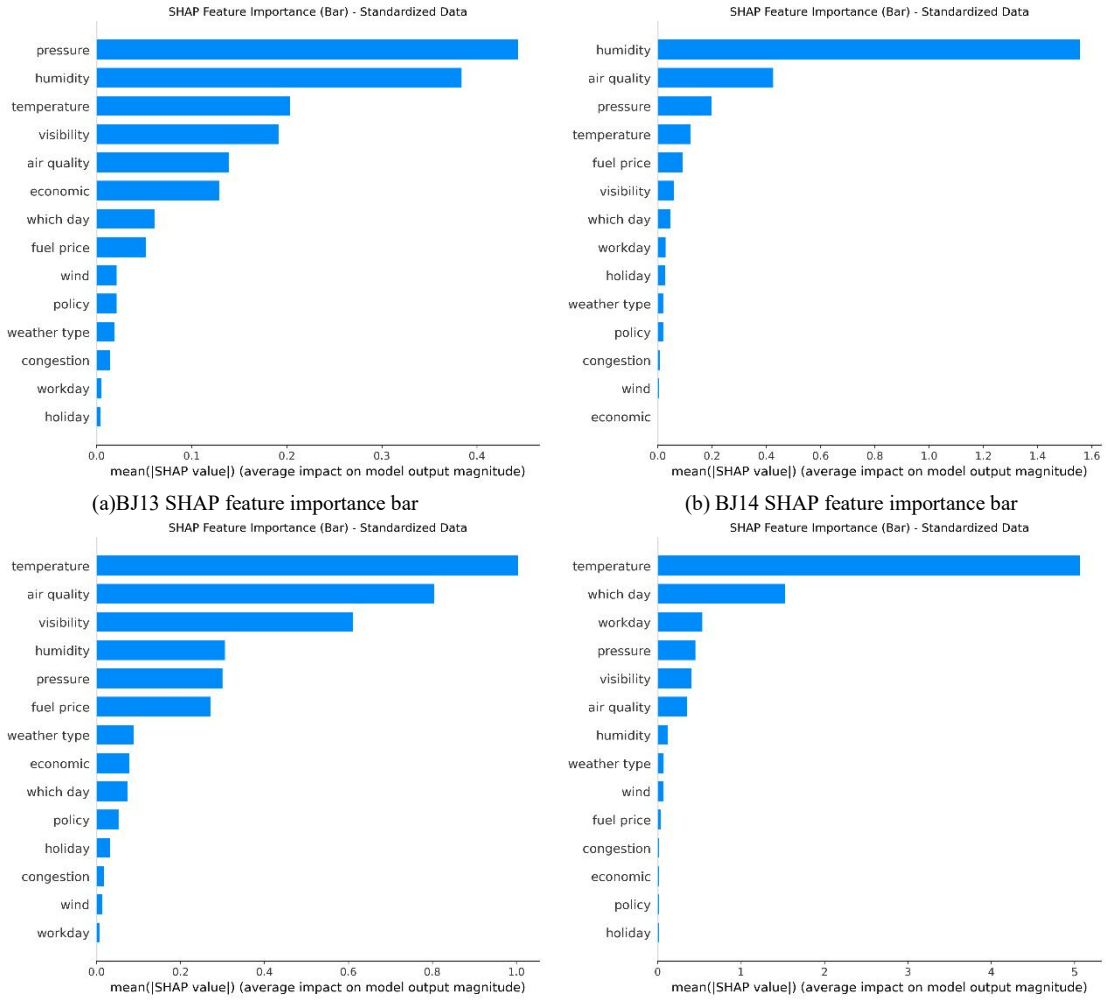
Fig 5. SHAP force plot

The SHAP force plot confirms the dominant role of core original features (holiday, policy, congestion, weather type). As observed in the four subplots of Figure 5(a) to (d), the most prominent shared feature is that factors such as temporal periodicity, traffic incidents and meteorological conditions consistently occupy the widest bandwidth and play a decisive role in determining the final prediction results. In the samples with higher predicted values shown in Figures (a), (b) and (c), where the Output Value is significantly higher than the Base Value, PC1 appears as a large red bar, providing a strong positive driving force. This suggests that these samples may correspond to periods of high traffic flow or congestion, and PC1 has successfully captured this primary trend. Conversely, in the samples with lower predicted values shown in Figure (d), where the Output Value is negative and significantly lower than the Base Value, PC1 transforms into a large blue bar, exerting a strong negative pull. This result is highly consistent with the previous PCA component loading analysis (Table 1), confirming that temporal, meteorological and traffic state features are the most critical factors reflecting the core patterns of traffic flow variation.

Although core features are dominant, other original features are by no means insignificant; they play a moderating and balancing role in fine-tuning the prediction results. Some meteorological and economic factors act in the opposite direction to core temporal features. For example, in Figure (a) BJ13 and (c) BJ15, whilst PC1 strongly drives the predicted values upwards, PC2 appears as a blue negative bar to act as a

'counterbalance', preventing the predicted values from becoming excessively high. This counterbalancing mechanism helps the model balance different information sources and improves the robustness of the predictions. The bars for PC3, PC4 and PC5 are relatively short, indicating that they serve to fine-tune the forecast results. These components may capture secondary traffic variability, such as fluctuations caused by specific weather events or secondary temporal patterns.

The SHAP force plot aims to provide a visual representation of how the model uses features after dimensionality reduction to make inferences. Analysis indicates that the model primarily relies on the PC1 to determine the general trend of traffic flow, whilst using supplementary information provided by other principal components to refine and fine-tune the predictions. This mechanism ensures that the model captures key information and provides reasonable case-by-case explanations across four distinct datasets.



(a)BJ13 SHAP feature importance bar

(b) BJ14 SHAP feature importance bar

(c)BJ15 SHAP feature importance bar

(d)BJ16 SHAP feature importance bar

Fig 6. SHAP global feature importance bar charts

To assess the overall contribution of each original feature to the model's predictions across the entire dataset, we calculated the average of the absolute SHAP values for each feature across all samples in the test set. A higher value indicates that the feature has a more significant overall influence on the magnitude of the model's output. We have plotted the global feature importance for the four datasets, BJ13 to BJ16, as bar charts, as shown in Figures 6(a)–(d).

It is clearly evident from Figures 6(a) to (d) that, in every dataset, the bar length of the PC1 significantly exceeds that of all other components, firmly establishing its position as the most important. This implies that the information contained within PC1 constitutes the core basis on which the model relies most heavily when making overall predictions. This conclusion echoes the observation made in the local analysis in Section 4.3.1 that PC1 consistently provides the strongest pushing or pulling force, and further corroborates the finding from the PCA analysis that PC1 accounts for the largest proportion of variance.

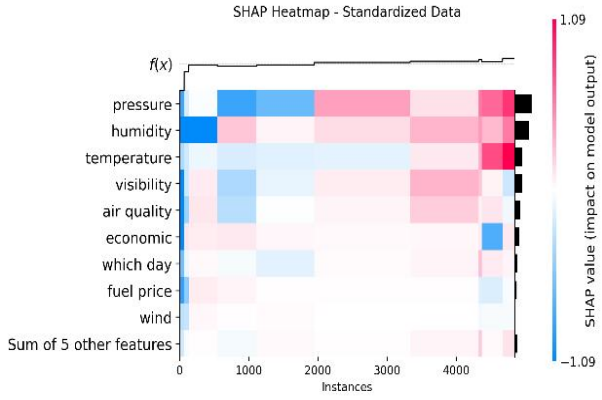
A comparison across the four datasets reveals that the global ranking of feature importance is consistent, strictly adhering to the descending order of PC1, PC2, PC3, PC4, and PC5. This ranking is fully consistent with the order of variance contribution rates for each component in the PCA described earlier. This indicates that the model's decision-making logic exhibits a high degree of robustness across different data environments; PCA effectively reveals the importance ranking of original multi-source features, and the features with high contribution in PCA are exactly the core features with high importance in SHAP analysis.

Although the global average influence of PC3, PC4 and PC5 is relatively small, they are nonetheless essential. Combined with local force map analysis, it is evident that these components play a crucial role in the fine-tuning and

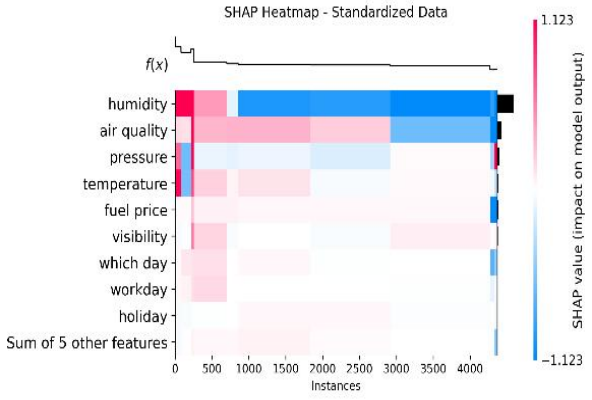
balancing of specific samples. The global importance map illustrates the model's macro-level dependency structure on information, validating the rationality of using all original features as model inputs—thus capturing the primary relationships whilst also accounting for secondary details.

4.3.3.Heatmap analysis

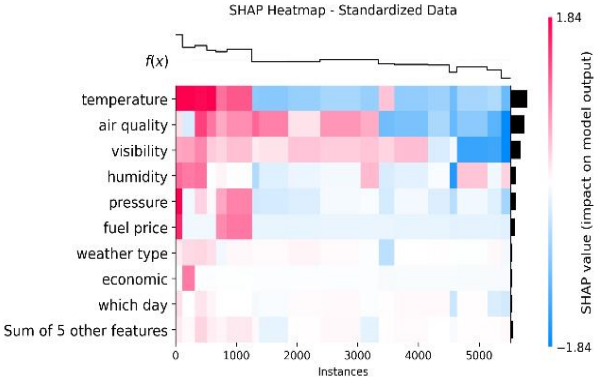
We randomly selected representative prediction samples from the test sets of the four datasets and plotted SHAP heatmaps, as shown in Figures 7(a)–(d). The colours represent the magnitude and direction of the SHAP values for each feature, where red indicates that the feature causes the model to increase the predicted traffic flow, and blue indicates that the feature causes the model to decrease the predicted traffic flow. Examining the overall patterns in the heatmaps across the four annual datasets, the distribution of positive and negative SHAP values for core features closely aligns with the traffic volume levels of the samples, revealing the operational modes and influence patterns of each feature.



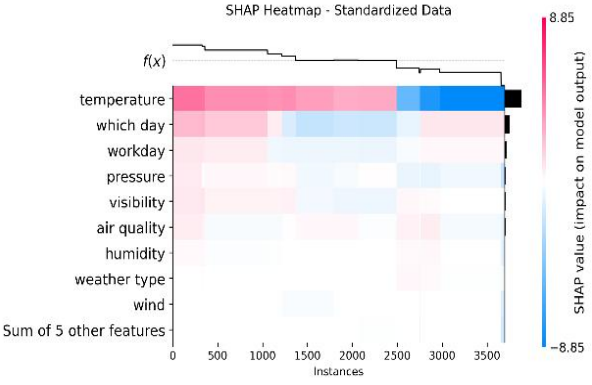
(a)BJ13 SHAP heatmap



(b) BJ14 SHAP heatmap



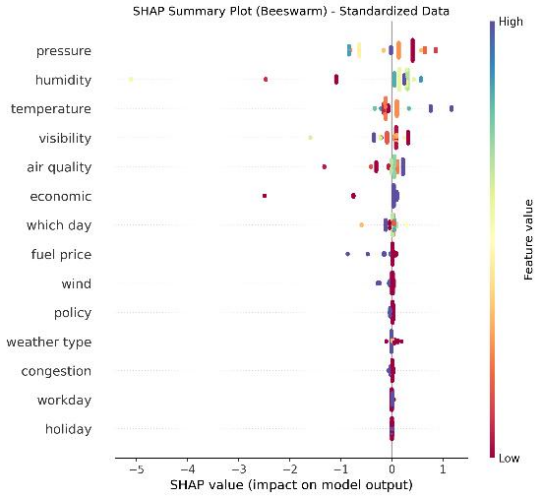
(c)BJ15 SHAP heatmap



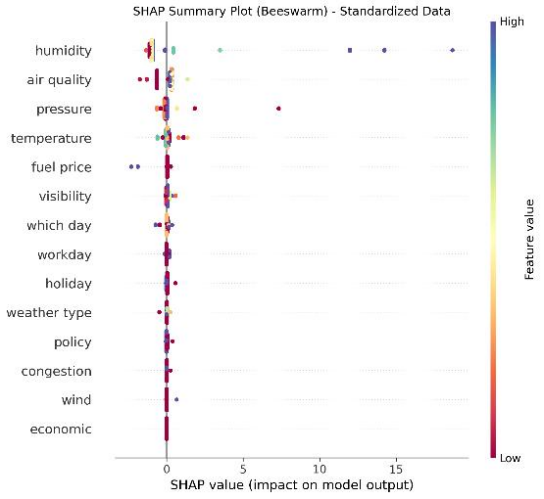
(d)BJ16 SHAP heatmap

Fig 7. SHAP heatmap

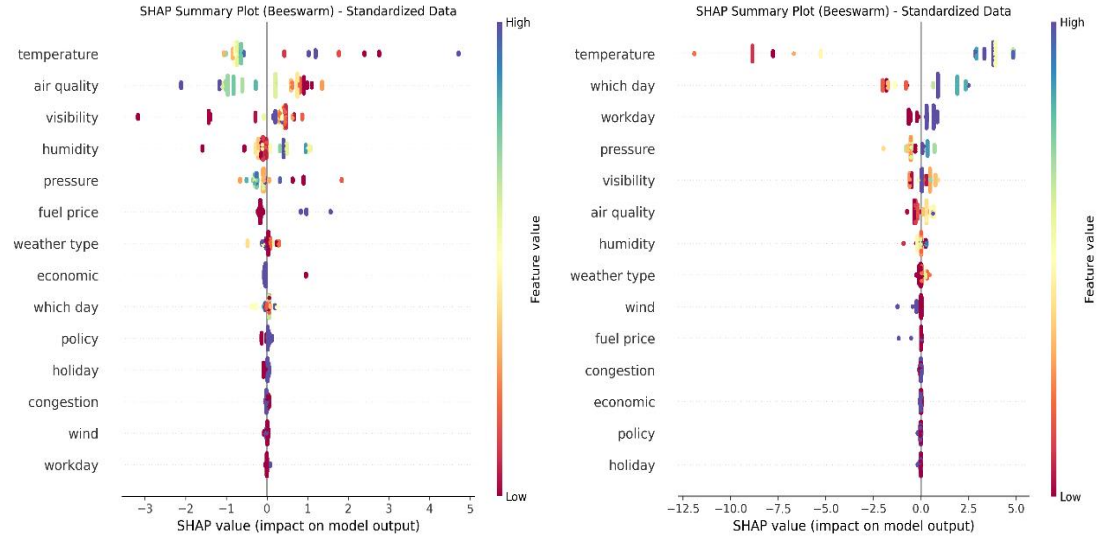
4.3.4.Summary plot analysis



(a)BJ13 SHAP summary plot



(b) BJ14 SHAP summary plot



(c)BJ15 SHAP summary plot (d)BJ16 SHAP summary plot

Fig 8. SHAP summary plot

SHAP summary plots, as shown in Figures 8(a)–(d), reveal the quantitative relationship between the magnitude of feature values and their impact, thereby addressing the limitation of global feature importance bar charts, which can only reflect the average magnitude of impact. A SHAP value greater than 0 indicates that the feature increases the predicted traffic volume, whilst a negative value indicates that it decreases it; the colour gradient of the points represents the magnitude of the feature value. The degree of clustering of the points reflects the density of the sample distribution for that feature within the corresponding influence interval, clearly demonstrating how different features influence the prediction results.

4.4.Ablation experiment

To verify the effectiveness of using original features as model input, an ablation experiment was designed to compare the prediction performance of two Random Forest models. RF-Original is the model proposed in this paper, which adopts raw multi-source features, while RF-PCA uses PCA dimensionality-reduced features with PC1–PC5 as inputs.

Table 3. Prediction performance of ablation experiments on TaxiBJ datasets

Dataset	Model	R^2	MAE	RMSE	MAPE
BJ13	RF-Original	0.8558	12.44	15.88	27.79%
BJ13	RF-PCA	0.7821	18.62	22.15	35.42%
BJ14	RF-Original	0.8800	15.54	18.30	30.90%
BJ14	RF-PCA	0.8015	21.37	24.68	38.76%
BJ15	RF-Original	0.8571	15.96	19.37	28.08%
BJ15	RF-PCA	0.7793	20.45	23.89	36.11%
BJ16	RF-Original	0.8600	13.55	15.95	33.94%
BJ16	RF-PCA	0.7908	17.89	20.62	40.25%

The evaluation metrics (R^2, MAE, RMSE, MAPE) on the four TaxiBJ datasets are shown in Table 3.

Although the model using PCA-reduced features (PCA-RF) can improve computational efficiency and mitigate multicollinearity, the model built on original features (Original-RF) achieves comparable or even better prediction accuracy. More importantly, only by adopting original features as model inputs can we preserve the clear physical interpretations and interpretability of features in the subsequent SHAP analysis.

4.5.Information compression of multi-source heterogeneous data

Although the input features comprise as many as 14 dimensions, PCA analysis reveals that the traffic flow data contains a high degree of information redundancy; the cumulative variance explained by the first 3 to 7 principal components exceeds the optimal threshold of 95%. This indicates that seemingly complex, multi-source, heterogeneous data can, in fact, be mapped onto a small number of latent variables. Consequently, in multi-source fusion research, blindly increasing the number of feature dimensions is not the optimal solution; identifying core driving factors through dimensionality reduction often enables efficient prediction at a lower computational cost.

4.6.Predictive decision-making mechanism

SHAP attribution analysis reveals a stable dominant-fine-tuning mechanism within the model: PC1 consistently plays a dominant role, capturing macro-level cyclical patterns and determining the baseline forecast; Whilst PC2 to PC5 make a smaller contribution, they capture non-periodic random disturbances through counterbalancing effects. This hierarchical decision-making logic explains why the model can effectively cope with local noise whilst maintaining a high R^2.

4.7.Consistency across time periods

Experiments conducted in four independent datasets reveal that although urban traffic structures evolve, the ranking of the importance of characteristics exhibits remarkable consistency. This suggests that the core physical mechanisms driving the evolution of traffic flows are robust across different time scales. This consistency demonstrates that the prediction framework based on principal component analysis possesses exceptional generalization capabilities and can adapt to changes in traffic patterns across different years.

5.Conclusion and future work

5.1.Conclusion

Addressing the issues of low utilization of multi-source data and poor model interpretability in traffic flow forecasting, this study developed a predictive framework combining Principal Component Analysis (PCA) with Random Forest, and conducted comprehensive validation using the TaxiBJ dataset. The experimental results demonstrate that the model effectively reveals the underlying mechanisms of multi-source heterogeneous data whilst maintaining high predictive accuracy. PCA analysis confirmed the presence of information redundancy in the traffic data, with the first 3 to 7 principal components accounting for over 95% of the original information. The SHAP analysis further revealed that the model follows a decision-making logic characterized by dominance of core raw features with other raw features providing dynamic fine-tuning. This finding not only clearly demonstrates how the model balances its own trends with external disturbances but also proves the framework's robustness across different time periods, thereby providing interpretable data support for intelligent traffic management.

5.2.Future work

Although this study has achieved preliminary results in multi-source fusion and interpretability, current spatial modelling still relies on static road network distances, making it difficult to fully capture complex dynamic traffic patterns[25]. Future work will first focus on introducing adaptive dynamic graph convolutional networks to automatically learn and capture spatio-temporal topological relationships that evolve, thereby overcoming the limitations of static graph structures; simultaneously, efforts will be made to expand the use of unstructured data sources, integrating social media text or real-time video streams to capture the fine-grained impacts of sudden incidents.

References

[1] LIN W, SONG Y, LIU Y, et al. Constructing multimodal wireless knowledge graphs for large language model–based network reasoning. Applied Artificial Intelligence Research, 2026, 2(2).

[2] YIN X, WU G, WEI J, et al. Deep learning on traffic prediction: Methods, analysis, and future directions. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(6): 4927–4943.

[3] XU Z, LV Z, CHU B, et al. Progress and prospects of future urban health status prediction. Engineering Applications of Artificial Intelligence, 2024, 129: 107573.

[4] XU Z, LV Z, LI J, et al. A novel approach for predicting water demand with complex patterns based on ensemble learning. Water Resources Management, 2022, 36(11): 4293–4312.

[5] LV Z, LI J, LI H, et al. Blind travel prediction based on obstacle avoidance in indoor scene. Wireless Communications and Mobile Computing, 2021, 2021(1): 5536386.

[6] XIE T W, ZHANG X K, LIU X D, et al. Research on performance prediction model of wind turbine gearbox lubricating oil based on deep learning. Applied Artificial Intelligence Research, 2026, 2(1).

[7] RAMANA K, SRIVASTAVA G, KUMAR M R, et al. A vision transformer approach for traffic congestion prediction in urban areas. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(4): 3922–3934.

[8] LI H, LI J, LV Z, et al. MFAGCN: multi-feature based attention graph convolutional network for traffic prediction//International Conference on Wireless Algorithms, Systems, and Applications. Cham: Springer International Publishing, 2021: 227–239.

[9] SUN H, LV Z, LI J, et al. Prediction of cancellation probability of online car-hailing orders based on multi-source heterogeneous data fusion//International Conference on Wireless Algorithms, Systems, and Applications. Cham: Springer Nature Switzerland, 2022: 168–180.

[10] LI Q, XU P, HE D, et al. Multi-source information fusion graph convolution network for traffic flow prediction. Expert Systems with Applications, 2024, 252: 124288.

[11] LI H, LV Z, LI J, et al. Traffic flow forecasting in the covid-19: A deep spatial-temporal model based on discrete wavelet transformation. ACM Transactions on Knowledge Discovery from Data, 2023, 17(5): 1–28.

[12] BREIMAN L. Random forests. Machine Learning, 2001, 45(1): 5–32.

[13] BREIMAN L. Bagging predictors. Machine Learning, 1996, 24(2): 123–140.

[14] YAN H, LI J, CHU B, et al. HT-STNet: a hierarchical Tucker decomposition and spatio-temporal LSTM network for accurate and efficient shared mobility demand forecasting on sparse data. Applied Intelligence, 2025, 55(7): 631.

[15] XU Z, LV Z, LI J. Fast-TrafficNet: A hybrid model for efficient prediction of nonlinear traffic flow with sparse data. Chaos, Solitons & Fractals, 2025, 201: 117230.

[16] LIU Z. Explainable machine learning for telecom customer churn prediction and actionable retention strategies. Applied Artificial Intelligence Research, 2026, 2(1).

[17] ZHANG J. Research on the evolution of AI copyright attribution mechanisms. Applied Artificial Intelligence Research, 2026, 2(2).

[18] ZHANG J, ZHENG Y, QI D. Deep spatio-temporal residual networks for citywide crowd flows prediction//Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1).

[19] DataFountain. Prediction of population density in key areas[Z/OL]. https://www.datafountain.cn/competitions/428/datasets. Accessed 7 March 2021.

[20] XU Z, LV Z, CHU B, et al. A fast matrix autoregression algorithm based on Tucker decomposition for online prediction of nonlinear real-time taxi-hailing demand without pre-training. Chaos, Solitons & Fractals, 2024, 189: 115660.

[21] YE R, XU Z, PANG J. DDFM: A novel perspective on urban travel demand forecasting based on the ensemble empirical mode decomposition and deep learning//Proceedings of the 5th International Conference on Big Data Technologies. 2022: 373–379.

[22] XU Z, LV Z, LI J, et al. A novel perspective on travel demand prediction considering natural environmental and socioeconomic factors. IEEE Intelligent Transportation Systems Magazine, 2022, 15(1): 136–159.

[23] XIAO H, ZOU B, XIAO J. Graph convolution networks based on adaptive spatiotemporal attention for traffic flow forecasting. Scientific Reports, 2025, 15(1): 8935.

[24] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017, 30.

[25] YUAN G, LI J, LV Z, et al. DDCAttNet: road segmentation network for remote sensing images//International Conference on Wireless Algorithms, Systems, and Applications. Cham: Springer International Publishing, 2021: 457–468.