Traffic
& Transportation

Regional Expressway Freight Volume Prediction Algorithm Based on Meteorological Information

Ning GAO[1], Yuanbo HONG[2], Junfei CHEN[3], Chonghao PANG[4]

Original Scientific Paper
Submitted: 7 Aug. 2022
Accepted: 16 Jan. 2023

[1] Corresponding author, Gao_peacewritng@163.com, School of Business, Hohai University

[2] 202121010099@mail.scut.edu.cn, School of Civil Engineering and Transportation, South China University of Technology

[3] chenjunfei@hhu.edu.cn, School of Business, Hohai University

[4] 15622245474@163.com, School of Civil Engineering and Transportation, South China University of Technology

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT

In the post-epidemic era, dynamic monitoring of expressway road freight volume is an important task. To accurately predict the daily freight volume of urban expressway, meteorological and other information are considered. Four commonly used algorithms, a random forest (RF), extreme gradient boosting (XGBoost), long short-term memory (LSTM) and K-nearest neighbour (KNN), are employed to predict freight volume based on expressway toll data sets, and a ridge regression method is used to fuse each algorithm. Nanjing and Suzhou in China are taken as a case study, using the meteorological data and freight volume data of the past week to predict the freight volume of the next day, next two days and three days. The performance of each algorithm is compared in terms of prediction accuracy and training time. The results show that in the forecast of freight volume in Nanjing, the overall prediction accuracies of the RF and XGBoost models are better; in the forecast of freight volume in Suzhou, the LSTM model has higher accuracy. The fusion forecasting method combines the advantages of each forecasting algorithm and presents the best results of forecasting the freight volumes in two cities.

KEYWORDS

road transportation; forecast of freight volume; machine learning; expressway; meteorological information.

1. INTRODUCTION AND LITERATURE REVIEW

In recent years, with the development of the Chinese economy and society and the advancement of expressway construction, transportation has progressively become a leading industry for the improvement of national comprehensive strength [1]. Expressway freight volume is the main indicator applied to measure the development of the regional transportation industry and reflects its economic development [2, 3]. In new economic conditions, especially in the post-epidemic era, dynamic monitoring of road freight volume is a prominent task [3, 4]. Accurate prediction of regional freight production not only provides a basis for formulating urban development plans but is also of great significance for rationally steering the allocation of public transportation resources [5, 6]. Accordingly, in the era of increasingly abundant basic data resources and the rapid development of deep learning and other technologies, the study of expressway freight volume forecasting methods has significant theoretical value and can provide ideas for different applications.

In recent years, with the development of the Chinese economy and society and the advancement of expressway construction, transportation has progressively become a leading industry for the improvement of national comprehensive strength [1]. Expressway freight volume is the main indicator applied to measure the development of the regional transportation industry and reflects its economic development [2, 3]. In new economic conditions, especially in the post-epidemic era, dynamic monitoring of road freight volume is a prominent task [3, 4]. Accurate prediction of regional freight production not only provides a basis for formulating urban development plans but is also of great significance for rationally steering the allocation of public transportation

resources [5, 6]. Accordingly, in the era of increasingly abundant basic data resources and the rapid development of deep learning and other technologies, the study of expressway freight volume forecasting methods has significant theoretical value and can provide ideas for different applications.

Numerous studies have demonstrated that adverse weather can decrease road adhesion and visibility, which in turn reduces the capacity of freight vehicles. Regarding the analysis of the impacts of different weather conditions on traffic flow parameters, rainfall is one of the most frequently analysed parameters [7–9]. Reduced road visibility caused by rainfall can make drivers more cautious while reducing speed and making journey times longer [10, 11]. Adverse weather conditions can also worsen road traffic and cause traffic congestion, leading to a range of adverse impacts, such as noise pollution and even traffic accidents [12, 13]. Hence, it is essential to integrate meteorological factors to forecast expressway freight generation trends.

The rest of the paper is organised as follows. Section 2 provides a literature review of relevant studies. Section 3 presents the data sources, the characteristics of the freight data set and the meteorological parameters employed. At the same time, the KNN, RF, LSTM, XGBoost and ridge regression algorithms principles are introduced. Section 4 proposes the experimental results and analysis of specific data sets, including the prediction accuracy and execution time of the five types of models. Section 5 concludes this paper.

2. LITERATURE REVIEW

There are abundant researches on the prediction of traffic volume related to transportation. Existing forecasting methods adopted by domestic and foreign scholars are mainly divided into classical forecasting methods based on statistical probability, machine learning forecasting methods based on intelligent interdisciplinarity and combined forecasting methods. Traditional forecasting methods generally include time series methods [14–16], regression analysis methods [17, 18] and gray forecasting methods [19]. The ARMA model [20, 21] is widely used in freight volume forecasting research as a classic time series forecasting method. The Gray-Markov forecasting model constructed by combining a gray prediction model and a Markov chain is more precise and efficient for prediction applications [22]. However, classical methods usually involve incompatible model assumptions and exhibit inferior performance in situations with complex traffic conditions [23].

Among the machine learning prediction methods, the commonly used traffic volume prediction models are the random forest (RF), extreme gradient boosting (XGBoost), k-nearest neighbour (KNN) and support vector machine regression (SVR). As a typical nonparametric method, the KNN model has received considerable attention. Many scholars have successfully applied a traditional KNN model to short-term traffic prediction [24–26]. Based on a traditional KNN model, Wu et al. [27] realised an enhanced model based on spatiotemporal information and argued that it achieved better performance than models that only employ temporal information. Filmon and Mecit [28] utilised an improved KNN algorithm to identify similar traffic patterns to forecast short-term passenger and freight traffic. Some authors applied an SVM to find the spatiotemporal correlation for traffic flow prediction. Feng et al. [29] utilised an adaptive multi-kernel SVM with spatial-temporal correlation (AMSVM-STC) for short-term traffic prediction. Lu and Gao [30] submitted an improved random forest regression (RFR) method to forecast and analyse railway freight traffic and found that its prediction had high accuracy, strong generalisation ability and great robustness. Chikaraishi et al. [31] used an XGBoost model to predict the traffic volume, and the results show that the model had high prediction accuracy and can improve computational efficiency. Deep learning (DL) technologies are being applied as mainstream nonparametric algorithms for feature extraction, pattern discovery and learning [32]. In terms of DL algorithms, long short-term memory (LSTM) neural network models are widely used for time series prediction [33, 34]. In addition, networks can draw on the mindset of human attention. Locations or features that require attention are assigned a higher weight. By incorporating an attention mechanism, a deep model can focus on fitting the attention features, and the feature extraction and fitting abilities of DL networks can be enhanced [32, 35].

Combined forecasting methods are achieved through the combination of several models to give full play to the advantages of each model in forecasting and obtain more accurate forecasting results [36, 37]. Qiao et al. [38] proposed a short-term traffic volume prediction method based on a convolutional neural network and long short-term memory (DCNN-LSTM), which achieved good prediction results. Vidas et al. [39] applied and compared four different machine learning algorithms (i.e. DT, RF, XGBoost and LSTM) for short-term travel time prediction. At present, few studies introduced meteorological information to forecast freight generation, and most are combined with machine learning algorithms to study the impact of meteorological information on expressway traffic flow forecasting. Soua et al. [40] used deep learning algorithms combined with weather data to forecast single-point traffic flow.

Literature review shows that machine learning models have been widely used in freight volume forecasting research. However, few studies take meteorological factors into account and analyse their impact on changes in freight volume and rarely utilise deep models with comparatively complex structures. In evaluating a model's effect, researchers only used the error value as the evaluation standard, the pros and cons of the model are not comprehensively analysed from the perspectives of the periodic prediction accuracy rate and the stability of the prediction effect of different step lengths. Furthermore, there is also a lack of comparative analysis between multiple models applied to the same data set.

Therefore, based on expressway toll data, this study utilises the historical data of freight volume in Nanjing and Suzhou, meteorological data, date and other time external characteristics as input parameters, and uses RF, XGBoost, LSTM and KNN models to forecast the freight volume of the target area. In addition, the prediction performance of the four models is comprehensively analysed from the aspects of average error, prediction accuracy and prediction effect stability under different prediction step sizes. Eventually, a fusion model based on ridge regression is proposed to comprehensively optimise the prediction ability of the four models.

The main contributions of this paper are as follows:

1) We compare the prediction effects of the RF, XGBoost, LSTM and KNN models under the input of the same data set and analyse their differences in prediction accuracy, periodic prediction accuracy, prediction stability with different step lengths, as well as training and prediction durations.
2) A fusion prediction model is proposed, which realises the fusion of multiclass models. It is proven that it has the highest prediction accuracy and the smallest prediction error.
3) Integrating meteorological factors and time series in forecasting of freight volume has significance in analysing the impact of weather on transportation and making transportation decisions.

3. METHODOLOGY

3.1 Data sources

A roadway network centre platform stores national expressway toll and portal information. Through statistics, the freight volume data of a research area can be obtained, which provides data support for the forecast of expressway freight generation. This research uses the expressway toll data of Jiangsu Province, and each row of data records the driving information of a motor vehicle, including vehicle type, the encrypted license plate index, entrance location, entrance time (accurate to seconds), exit location and exit time (accurate to seconds).

We performed a one-hot code for seven types of weather conditions (sunny, cloudy, overcast, hazy, light rain, moderate rain, heavy rain) for each day in the study area, and normalised wind level and temperature variables, which were combined with the normalised time, date, number of weeks, average freight volume and trip data of 6 categories of vehicles to form an input vector. Simultaneous input of data for consecutive days forms an input matrix. Furthermore, the training set and the testing set are divided on this data set, and the K-nearest neighbour algorithm is used to predict the freight volume.

d, d_h and d_p are used to represent the days of data coverage, the days of historical data and the days of prediction, respectively, so the calculation formula of the input sample number N_{in} is:

$$N_{in} = d - d_h - d_p \tag{1}$$

The input matrix X and the true value matrix Y can be expressed as:

$$X = (x^{(1)}, x^{(2)}, ..., x^{(s)} \tag{2}$$

$$Y = (y_1, y_2, ...y_N)^T \tag{3}$$

where $x^{(s)}$ represents the s-th input feature. Then, the training set can be expressed as:

$$T = (X, Y) = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\} \tag{4}$$

The training process of each model seeks to find the optimal mapping from X to Y in the training set to minimise the error between the predicted value and the true value. This mapping is expressed as:

$$\hat{Y} = \phi_{T\ S}(X) \tag{5}$$

where ϕ represents the mapping from input to output and T represents the training set.

The notations utilised in this paper are described in *Table 1*.

Table 1 – Notations and descriptions

Notations	Descriptions
d	The days of data coverage (364 days)
d_h	The days of historical data (7 days)
d_p	The days of prediction (1 day, 2 days and 3 days)
N_{in}	The number of the input sample
$X=(x^{(1)}, x^{(2)},\ldots, x^{(s)})$	Input matrix
$Y=(y_1, y_2,\ldots, y_N)^T$	True value matrix
$T=(X,Y)$	Training data
$\hat{Y}=\phi_{T,S}(X)$	Forecasted data from models

3.2 Freight volume time series

Using the toll data of the expressway network centre platform for statistics, a set of time series containing the average freight volume and operating trip data of six types of vehicles is obtained. The formula for calculating the freight volume is as follows:

$$AHCTV=\frac{\sum_{K}(w_h^k - w_{self}^k)}{K} \tag{6}$$

where *AHCTV* represents the average load of each freight vehicle, w_h^k represents the total weight of the *k*-th vehicle and w_{self}^k represents the self-weight of the *k*-th vehicle. *K* represents the total number of vehicles. According to the value recommended by the Ministry of Transport Planning and Research Institute, the self-weight of different types of freight vehicles is shown in *Table 2*.

Table 2 – Self-weight of freight vehicles by model

Truck by model	Cargo 1	Cargo 2	Cargo 3	Cargo 4	Cargo 5	Cargo 6
Self-weight [t]	2.036	5.728	10.547	13.565	15.290	16.298

This study takes Nanjing and Suzhou in China as the research objects, selects the charging data of the two cities in 2021 and counts the average freight volume and running data of six types of vehicles from 00:00 to 23:59 throughout the day. According to the statistics, the daily freight volumes of Nanjing and Suzhou in 2021 are shown in *Figure 1*. As seen from the figure, the cyclical changes in the freight volume data in Nanjing and Suzhou in 2021 are drastic and show a significant downward trend on weekends. In addition, during the Spring Festival and various holidays, the freight volumes of the two cities are at their lowest points of the year.

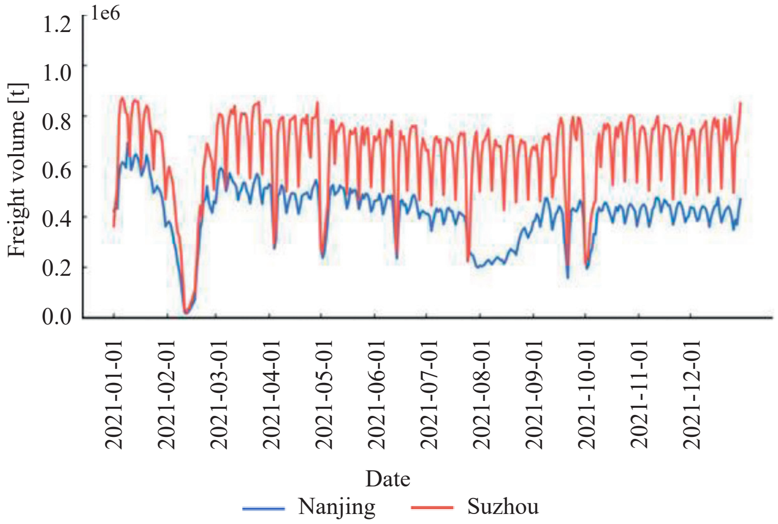


Figure 1 – Freight volumes of Nanjing and Suzhou in 2021

3.3 Quantification of meteorological factors

The American "Handbook of Expressway Capacity" expounds on the effect of rainfall on expressway operating speed and capacity [41], indicating that light rain will reduce the free-flow vehicle speed by approximately 1.9 km/h and heavy rain will reduce the free-flow vehicle speed by 4.8–6.4 km/h. The impact of rainfall intensity on traffic capacity is shown in *Table 3*.

Table 3 – Changes in the influence of rainfall intensity on traffic capacity

Weather variable	Rainfall intensity [inch/h]	Capacity reduction ratio (%)
Rain	0–0.1	1.2–3.4
	0.1–0.25	5.7–10.1
	>0.25	10.7–17.7

The air temperature primarily affects the temperature of the engine compartment temperature of a vehicle and the fatigue degree of a driver. The road temperature will affect the driving conditions of the vehicle tires. In summer, the high road temperature and the increased friction between the tires and the road lead to unstable tire pressure and the aging of tire rubber, which simply causes the vehicle to lose control. When the average air temperature along an expressway exceeds 33°C and the average road surface temperature exceeds 55°C, the incidence of traffic crashes increases. In winter, the temperature is low, the road surface is prone to freezing, the adhesion coefficient decreases and the tires of motor vehicles slip, which is similar to the decrease in road capacity caused by heavy rainfall.

This study primarily selects air temperature, weather conditions (sunny, cloudy, overcast, haze, light rain, moderate rain, heavy rain) and wind level as research factors to study and forecast the freight volume in Nanjing and Suzhou. The weather conditions during the year are illustrated in *Tables 4–6*.

Table 4 – Weather conditions in Nanjing and Suzhou in 2021

Weather	Sunny	Cloudy	Overcast	Haze	Light rain	Moderate rain	Heavy rain
Nanjing (day)	93	135	74	3	41	12	7
Suzhou (day)	96	110	81	1	54	19	4

Table 5 – Nanjing and Suzhou wind levels in 2021

Wind level	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Nanjing (day)	4	36	170	134	20	1	0
Suzhou (day)	1	25	163	156	18	1	1

Table 6 – Temperatures in Nanjing and Suzhou in 2021

Temperature [°C]	<0	[0,5)	[5,10)	[10,15)	[15,20)	[20,25)	[25,30)	>30
Nanjing (day)	7	22	52	72	54	38	113	7
Suzhou (day)	6	16	55	74	49	37	107	7

3.4 Methodological framework

K-nearest neighbour algorithm

KNN is a mode recognition algorithm based on an instance, whose basic principle is to correspond to any n dimensional input vector in the feature space and then extract the nearest data category label or predictive value of the nearest data as the output.

Using n as the capacity of the sample data set, K is the only hyperparameter in the KNN algorithm. Generally, its value is a small, odd number, whose range is often $(1,\sqrt{n})$.

In the feature space, the distance between two instance points reflects the similarity between them. In KNN, the distance measurement function is defined as follows:

$$L_p(x_i,x_j)=\left(\sum_{i=1}^{n}|\,x_i^{(l)}-x_j^{(l)}\,|^p\right)^{\frac{1}{p}} \tag{7}$$

where $x_i,x_j \in R^n$. When p=2, the distance between points is the Euclidean distance, which is the most commonly used distance function.

To be used in regression, the KNN algorithm quantifies labels corresponding to data vectors, applying an average value method or weighted average value method to attain the predicted value of the test point. The weighted average value formula is defined as:

$$y_0=\frac{1}{k}\sum_{i}^{k}y_i\cdot \mathrm{dist}(\vec{x},\vec{t}) \tag{8}$$

where *dist* is the distance function between vectors and $\vec{t}$ is the data point classified by position.

For high-dimensional input, the radial basis function is often used to assign a weight to K near-neighbouring points and the calculation method is:

$$W_K(x_0,x_k)=\frac{1}{K}\ \exp\left\{-\frac{\left\|x_k-x_0\right\|^2}{2K}\right\} \tag{9}$$

During KNN algorithm training, the K value can be adjusted to improve the fitting ability of the algorithm. The use of a KD tree to store data for search can solve the problem of slow speed caused by an excessive number of samples.

Random forest algorithm

An RF is a representative bagging integrated algorithm. Based on decision trees as a base learner, random selection features are introduced during the training process.

The principle is to select K attributes from all attributes randomly and then choose the best segmentation attribute as a node to establish a CART decision tree. For the quality of the division attributes and the division point, we use the average impurity $G(x_i, v_{ij})$ of each child node to measure, and the calculation formula is as follows:

$$G(x_i,v_{ij})=\frac{n_{left}}{N_s}H(X_{left})+\frac{n_{right}}{N_s}H(X_{right}) \tag{10}$$

where x_v, v_{ij} are a certain kind of segmentation attribute and its division value, n_{left}, n_{right}, N_s are the number of training samples of the left and right nodes after the division and the total number of samples of the current node, X_{left}, X_{right} is the collection of left and right child node training samples and $H(X)$is the impurity function of nodes. For regression, the mean square error (MSE) and mean absolute error (MAE) are usually used.

The training process of the decision tree is equivalent to an optimisation problem at a certain node, that is, to find the minimum segmentation attribute and dividing point of G:

$$(x^*,v^*)=\arg\min_{x,v}G(x_i,v_{ij}) \tag{11}$$

The RF adopts multiple different decision trees to increase the robustness and stability of the final model prediction results. Decision tree samples and bootstrapping are parameters used to control random sampling technology, indicating that training sample set D is repeatedly sampled N times and the parameters are returned each time. The probability selected for each sample is $1/N$, so the probability of the selection after N time sample is:

$$(1-\left(1-\frac{1}{N}\right)^N \Rightarrow (1-\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^N=\left(1-\frac{1}{e}\right)^N \cong 0.633 \tag{12}$$

By satisfying the randomness of characteristics and samples, the RF algorithm compensates for the shortcomings of decision trees in weak generalisation. During training, the number of decision trees and maximum depth can be adjusted to improve the fitting ability of the algorithm.

Extreme gradient boosting

XGBoost is a boosting integrated learning algorithm whose base learner is a CART or linear classifier (gblinear) and the learning of the base learner is serial. The CART regression tree's additional model can be expressed as:

$$\hat{y}_i = \phi(x) = \sum_{i=1}^{K} f_k(X),\ f_k \in F \tag{13}$$

where $F = \left\{ f(x) = \omega_{q(x)} \right\} (q : R^m \to T, \omega \in R^T)$ is the function space comprising all regression trees. f_k corresponds to the leaf node score of an independent tree structure *q*. *q* is a function reflecting the sample point to the leaf point, and *T*, ω_i are the number and score of the leaf nodes. Based on regularisation thought, for a given data set $\left\{ (x_i, y_i), 1 \le i \le n \right\}$, it can be useful to learn the model by optimising the following target functions:

$$\min_{fk} L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{14}$$

where $\Omega(f) = \gamma T + \frac{1}{2} \lambda \, || \, \omega \, ||^2 \cdot l$. *l* is the loss function and Ω indicates the complexity of the tree. *γ*, *λ* are the regularisation coefficients, which belong to the hyperparameter.

By finding a partial derivative for the loss function, the target function can be inferred as:

$$\tilde{L}_q^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{15}$$

$\tilde{L}_q^*$ is only related to the structure of the tree. It has nothing to do with the leaf nodes. The better the tree structure *q* is, the smaller $\tilde{L}_q^*$ is.

The maximum depth, contraction and characteristic subspace sampling of the tree are restricted to relieve model overfitting.

Long short-term memory

LSTM is a deformation structure of a recurrent neural network (RNN). It adds memory units to each hidden layer, making the memory information on the time series controlled, thereby overcoming the long-term dependence of an RNN.

LSTM controls the deposition or addition of information through a gate to achieve the function of forgetting or memory. A forget gate is a *sigmoid* function belonging to h_{t-1}, the output from the last former unit and the x_t input from this unit. Each item C_{t-1} ranged in [0,1] is used to control the forget level of the last former unit, which is shown as follows:

$$f_t = \sigma(W_f \cdot \left[h_{t-1}, x_t \right] + b_f) \tag{16}$$

The input gate contains the output i_t of the *sigmoid* activation function and the output $\tilde{C}_t$ of the *tanh* function, which controls what values can be used to update values and create new values, expressed as:

$$i_t = \sigma(W_i \cdot \left[h_{t-1}, x_t \right] + b_i) \tag{17}$$

$$\tilde{C}_t = \tanh(W_c \cdot \left[h_{t-1}, x_t \right] + b_c) \tag{18}$$

Based on this situation, we can update the cell status of this unit:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{19}$$

The output gate controls how much the cell status of this layer is filtered. First, the *sigmoid* activation function is used to obtain o_t with a range of [0,1]. Furthermore, each pair of cell states C_t is multiplied after the *tanh* function treatment. Finally, the model output h_t can be obtained as follows:

$$o_t = \sigma(W_o \cdot \left[h_{t-1}, x_t \right] + b_o) \tag{20}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{21}$$

Aiming at minimising the error of real values and prediction values, the loss function is built up:

$$\min J(\theta) = \sum_{t=1}^{T} \text{loss}(\hat{y}^{(t)}, y^{(t)}) \tag{22}$$

LSTM performs parameter training with the use of a backpropagation algorithm by constantly adjusting weight and bias, which leads to the predicted value closer to the real value. The model framework of LSTM is shown in *Figure 2*.

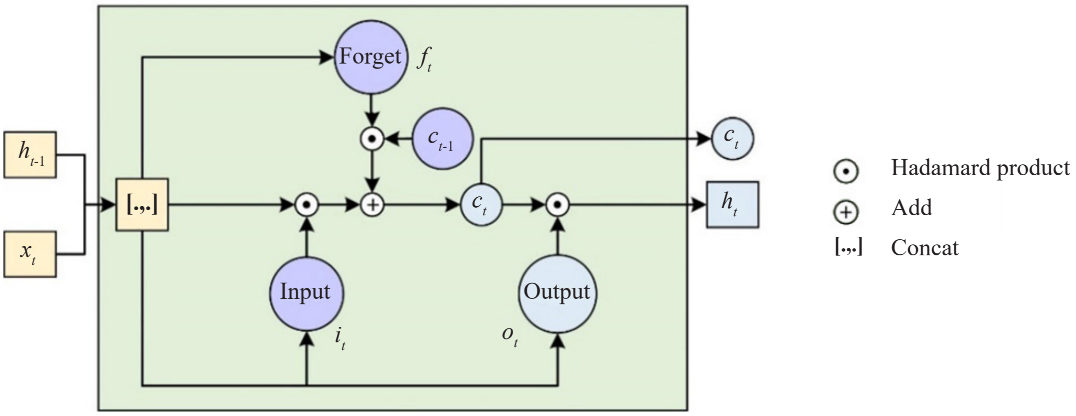


Figure 2 – The model framework of LSTM

Ridge regression fusion model

Ridge regression is a regression model used to deal with situations where the number of characteristics is more than samples or multicollinearity between characteristics. It transforms the process of solving the regression coefficient ω into an optimisation problem with conditions and then solves it using the minimum daily method.

The target function of linear regression is:

$$J(\omega)=\sum(y-X\omega)^2 \tag{23}$$

The ridge regression model adds a paradigm punishment item based on it. It is the regularisation coefficient. The expression of the loss function becomes:

$$J(\omega)=\sum(y-X\omega)^2+\lambda\,||\,\omega\,||_2^2 \tag{24}$$

Inferred from that, the regression coefficient is:

$$\omega=(X^T X+\lambda I)^{-1}X_y^T \tag{25}$$

The addition of L_2's punishment items make $(X^T X+\lambda I)$ full rank to ensure that it is reversible. λ can avoid the impact caused by precise correlation. Adjusting λ can control the offset of the parameter vector ω. The larger λ is, the more difficult it is for the model to be affected by multicollinearity. Nonetheless, when it is too large, the estimate ω has a large offset and cannot fit the real appearance of the data. Therefore, we need to find the best λ value of the model, the common methods include the $\omega-\lambda$ ridge curve and cross-validation.

Ridge regression abandons the unbiasedness of the least squares method and obtains the regression coefficient at the cost of losing part of the information and reducing accuracy. Moreover, it is a more actual and reliable method than linear regression, which can alleviate overfitting problems.

Previous researches have shown that a proper assembly framework can take advantage of a set of learning algorithms and obtain better prediction than any single method [42, 43]. It is an effective way to ensemble multiple prediction methods in order to integrate various advantages from selected methods set [44].

Instead of selecting a single learning method for prediction, an assemble method based on ridge regression is developed to combine the strengths of multiple predictive modelling methodologies. Ridge regression has obvious advantages in solving the linear regression overfitting problem as well as ill-conditioned data, so we first use four commonly used machine learning models (random forest (RF), extreme gradient boosting (XGBoost), long short-term memory (LSTM) and K-nearest neighbour (KNN)) to find the nonlinear mapping of X–$\hat{Y}$, and then use the ridge regression method to adjust the weights for linear fitting, so as to improve the prediction accuracy. More importantly, the experimental results also verify this point, and the fusion model combines the advantages of each prediction algorithm and can be used to obtain superior overall predictive performance. The framework of this prediction method is shown in *Figure 3*.

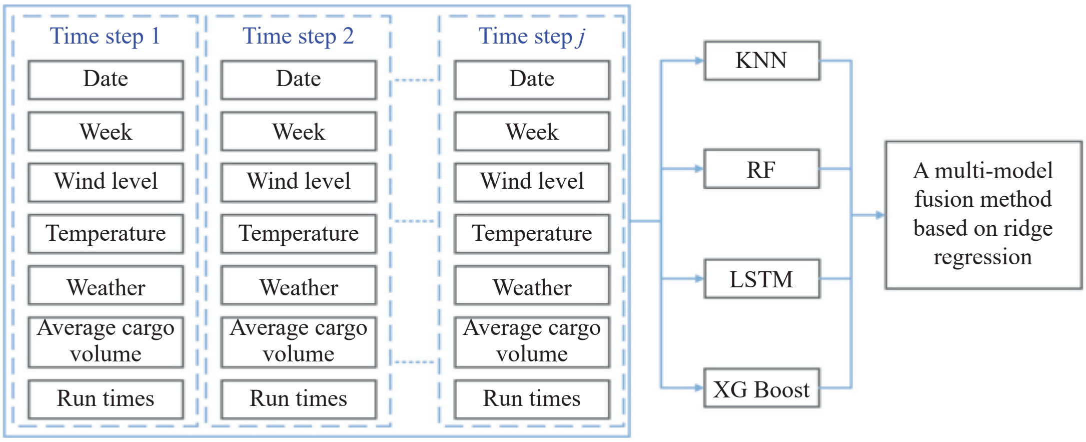


Figure 3 – Framework of the fusion prediction model based on ridge regression

3.5 Model prediction performance verification index

The indexes used to evaluate the prediction accuracy include the mean absolute error (MAE), mean absolute percentage error (MAPE) and mean square error (MSE).

The MAE avoids the positive and negative disappearance of errors, which can better reflect the actual situation of prediction errors. The calculation formula is:

$$\mathrm{MAE} = \frac{1}{N_{pr}} \sum_{i=1}^{N_{pr}} \left| y_i - \hat{y}_i \right| \tag{26}$$

The MAPE is the ratio of the absolute error to the real value. It is used to reflect the reliability of different measurement results. The calculation formula is:

$$\mathrm{MAPE} = \frac{100\%}{N_{pr}} \sum_{i=1}^{N_{pr}} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{27}$$

The MSE can comprehensively reflect the difference between the prediction value and the real value. The calculation formula is:

$$\mathrm{MSE} = \frac{1}{N_{pr}} \sum_{i=1}^{N_{pr}} (y_i - \hat{y}_i)^2 \tag{28}$$

In *formulas 26–28*, N_{pr} is the number of prediction samples, y_i is the real value of the i-th time series and $\hat{y}_i$ is the predicted value of the i-th time series.

4. RESULTS

In the experiment, each model uses the same time step for prediction and output, the input step is 7 days and the prediction step is 1 day, 2 days and 3 days. We adopt the same input in the KNN model, RF model, LSTM model and XGBoost model, and the input to the ridge regression model is the output results of the first four models. For the data set division, the data from the first 255 days (1 January 2021 – 12 September 2021) are used as training samples, and the data from the last 109 days (13 September 2021 – 30 December 2021) are used as a test sample.

4.1 Analysis of prediction accuracy

The errors of each model under different prediction step sizes are shown in *Tables 7–9* and *Figures 4 and 5*. It can be seen that the prediction error of each model generally increases with increasing prediction step size. For each step size, the errors of the RF, XGBoost and fusion models are lower than those of the KNN and LSTM models when predicting the freight volume in Nanjing, whereas the LSTM model has the lowest error when predicting the freight volume in Suzhou.

Table 7 – Mean square error of different models

	Prediction step (day)	MSE (E+09)				
		KNN	RF	LSTM	XGBoost	Fusion model
Nanjing	1	3.2685	1.1680	2.1626	1.1181	0.3889
	2	3.6615	2.3021	3.8370	1.7573	0.3613
	3	3.3824	2.1378	4.7663	1.8286	0.5087
Suzhou	1	12.2681	7.2708	6.5559	7.6312	3.0605
	2	12.8291	14.5936	12.6203	14.0974	8.2204
	3	11.3099	20.0618	12.4388	16.3079	7.0805

Table 8 – Mean absolute error of different models

	Prediction step (day)	MAE (E+04)				
		KNN	RF	LSTM	XGBoost	Fusion model
Nanjing	1	4.0653	2.2945	3.6242	2.2169	1.4435
	2	4.2002	2.9606	4.6809	2.6151	1.5046
	3	4.0614	3.1079	5.3513	2.9107	1.7554
Suzhou	1	6.4989	5.0176	4.8117	5.5390	4.2763
	2	6.7899	9.7297	6.8172	9.3463	6.3756
	3	6.7492	12.2902	7.4637	10.6173	6.2023

Table 9 – Mean absolute percentage error of different models

	Prediction step (day)	MAPE				
		KNN	RF	LSTM	XGBoost	Fusion model
Nanjing	1	0.1245	0.0695	0.1020	0.0664	0.0372
	2	0.1277	0.0923	0.1332	0.0808	0.0377
	3	0.1156	0.0888	0.1448	0.0836	0.0453
Suzhou	1	0.1524	0.1093 0.1789	0.1052	0.1152	0.0693
	2	0.1559	0.1789	0.1482	0.1726	0.1017
	3	0.1386	0.2036	0.1454	0.1778	0.1042

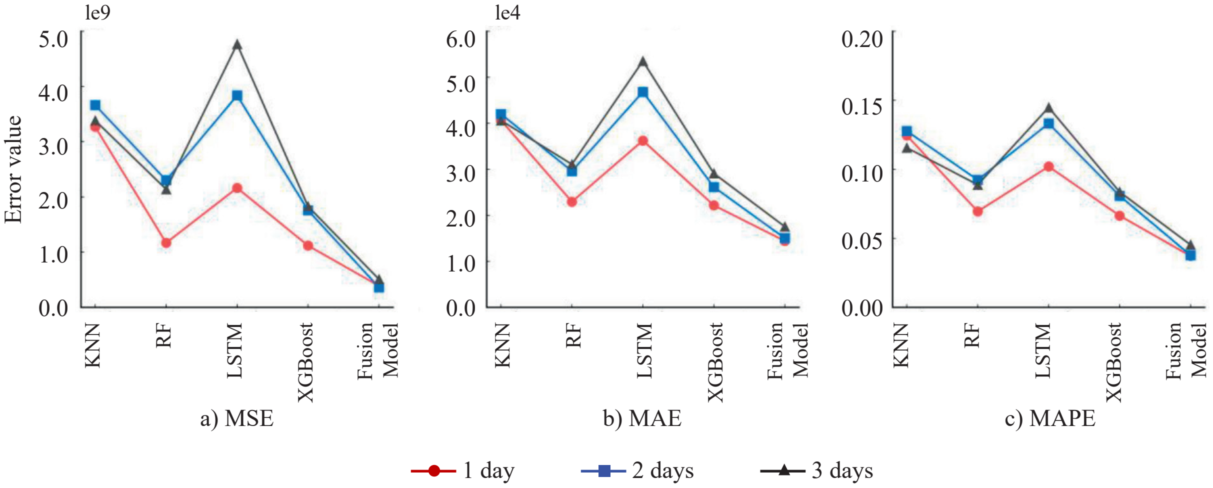


Figure 4 – Error curves of each model under different prediction steps in Nanjing

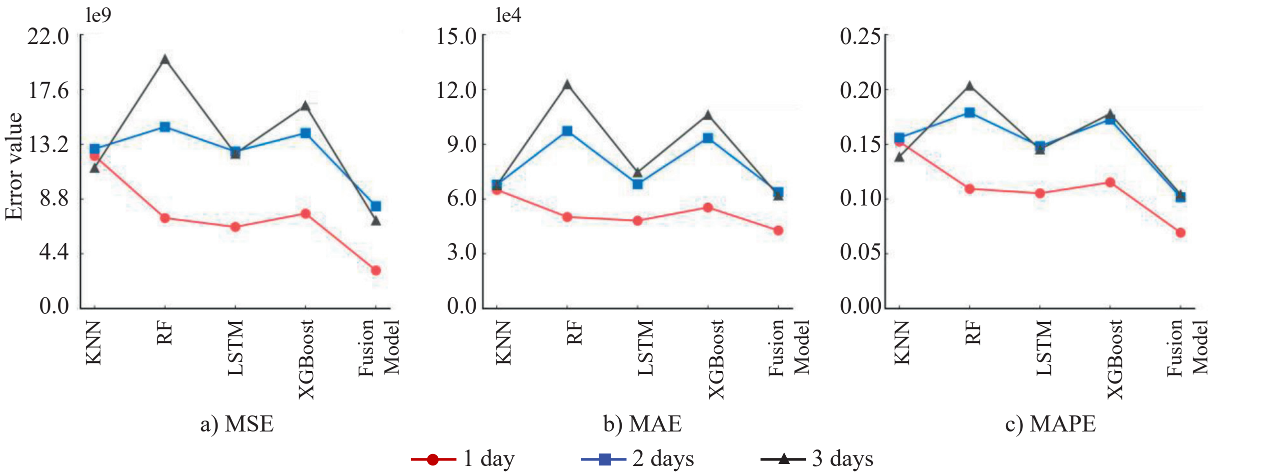


Figure 5 – Error curves of each model under different prediction steps in Suzhou

To further compare and analyse the prediction effects of the four models, we used box plots and the cumulative distribution function (CDF) to display the prediction errors. It can be seen from *Figures 6 and 7* that the median and upper quartile of the MAPE of each model generally increase slightly with increasing prediction steps, and the maximum upper quartile is lower than 25%, which indicated the stability of the prediction performance of each model under different step lengths. The CDF curves in *Figures 8 and 9* show the cumulative distribution of errors in each model under the same step length. The abscissa of a point on the curve represents the value of the MAPE, and the ordinate represents the proportion of samples whose MAPE is lower than the abscissa of the point. From the figures, we can see that the fusion, RF and XGBoost models perform better than the KNN and LSTM models at the same time length for the fitting of Nanjing freight data. However, for the fitting of Suzhou freight data, the fusion and LSTM model performances are better.

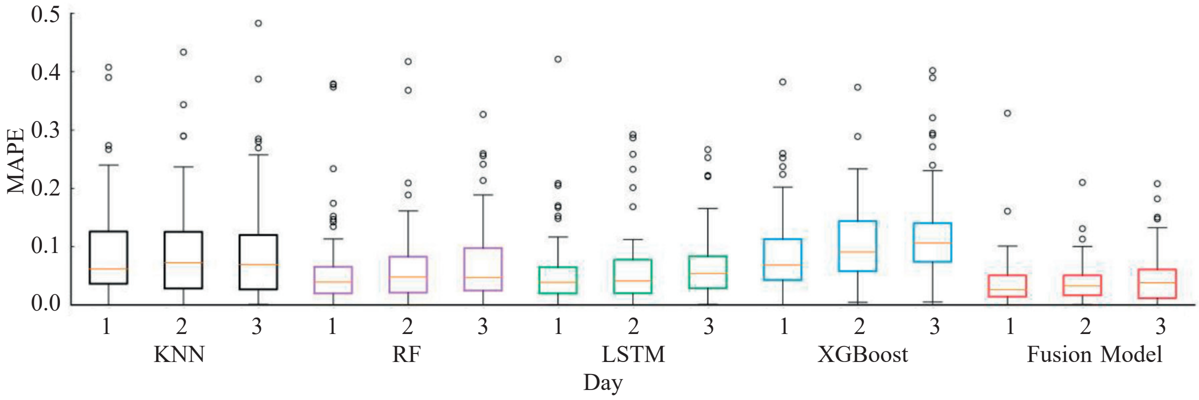


Figure 6 – Prediction error box diagram of each model under different prediction steps in Nanjing

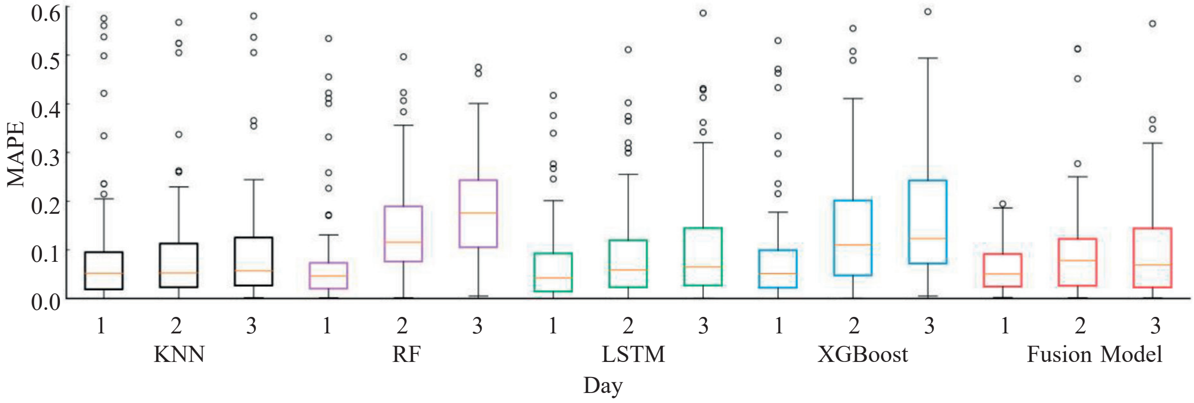
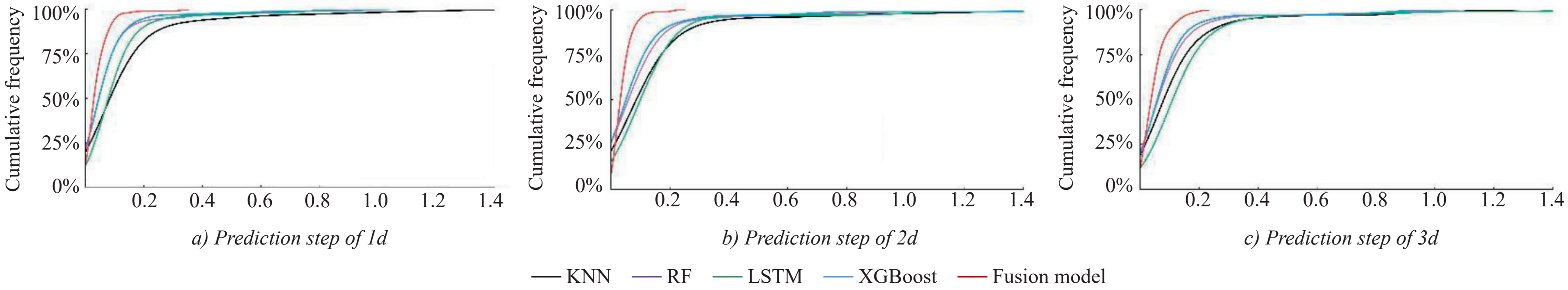


Figure 7 – Prediction error box diagram of each model under different prediction steps in Suzhou

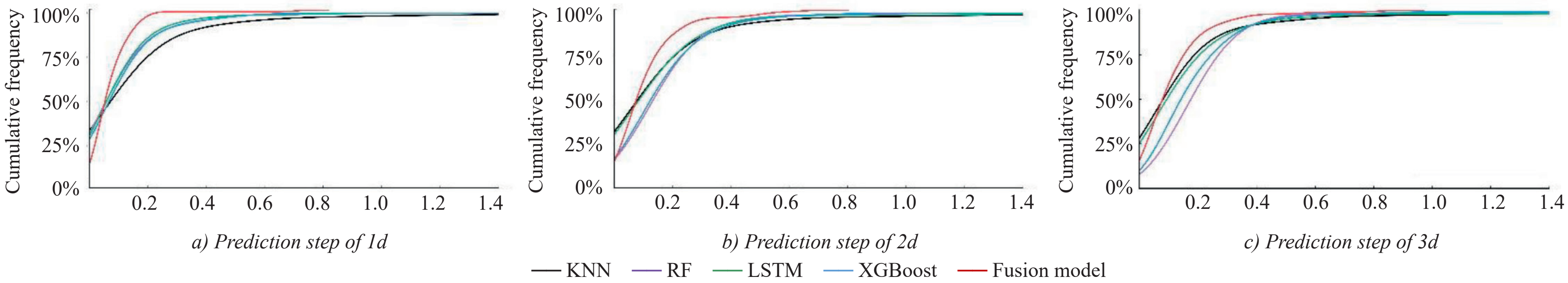


a) Prediction step of 1d

b) Prediction step of 2d

c) Prediction step of 3d

Figure 8 – CDF curve of each model under the same prediction step in Nanjing



a) Prediction step of 1d

b) Prediction step of 2d

c) Prediction step of 3d

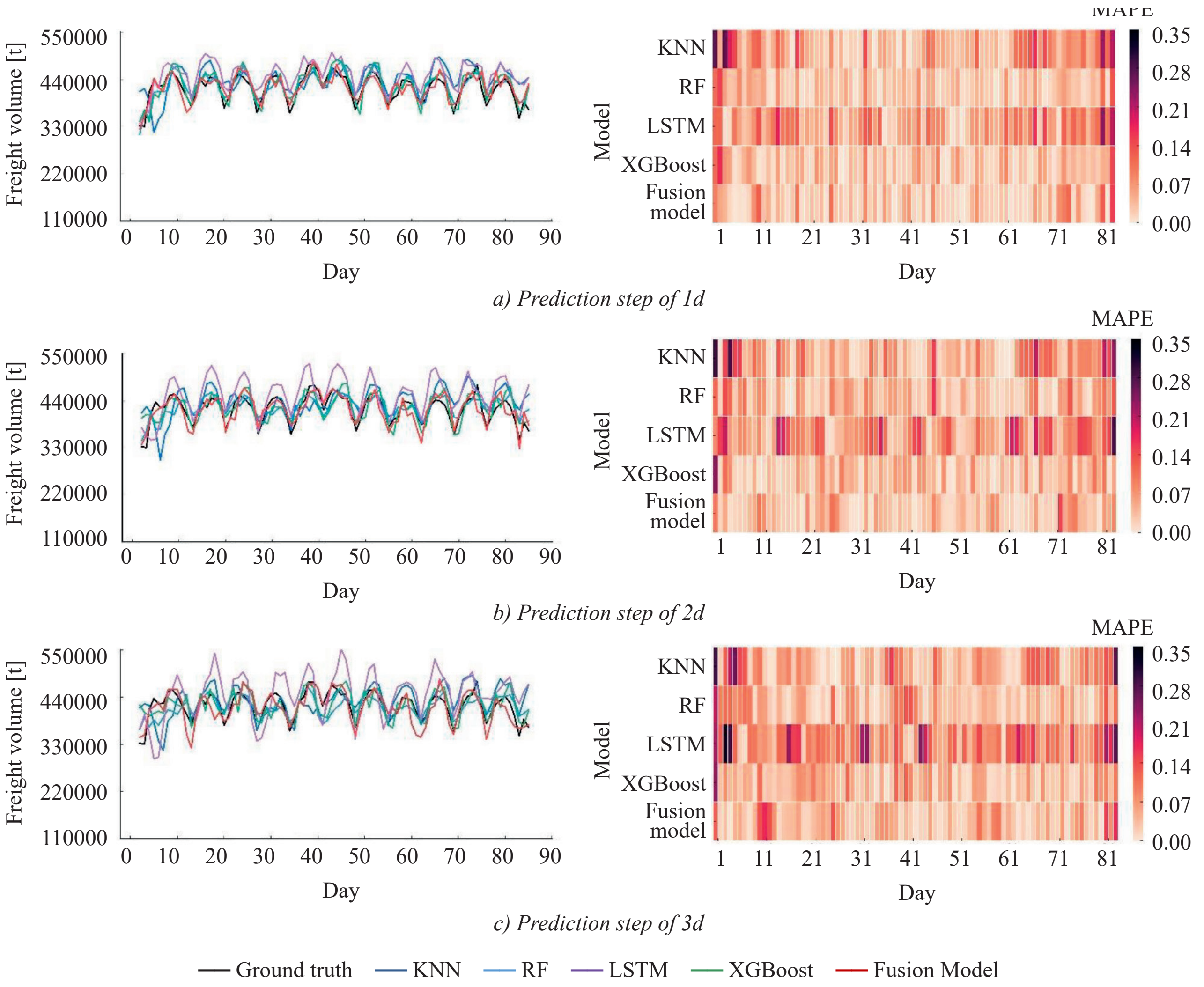
Figure 9 – CDF curve of each model under the same prediction step in Suzhou

We selected 84-day data of Nanjing and 86-day data of Suzhou with obvious periodic changes, and the prediction errors of each model under three step sizes were visualised. The results are shown in *Figures 10 and 11*. It can be seen that the prediction error of the LSTM model is relatively large in the forecast of freight volume in Nanjing, and the prediction abilities of the fusion, RF and XGBoost models are better. In contrast, the forecast results of freight volume in Suzhou reveal that the fusion and LSTM models have the best fitting effects, indicating that the LSTM model has a better effect on the data set with significant periodic changes.

To sum up, the RF, XGBoost, LSTM and fusion models based on ridge regression may be selected for prediction with a short prediction step, while the fusion model is the best choice for long-step prediction. Moreover, the LSTM and fusion models can be selected for prediction when periodic data set changes evidently.

4.2 Analysis of execution time

The training duration and prediction duration of each model are shown in *Table 10*, in which the LSTM model performed 100 rounds of training. During training, the KNN model only builds a fast search structure, while the other models store parameters, which results in a shorter training time for the KNN model. During prediction, the KNN model needs to calculate or find similar samples in the training set, leading to a longer prediction execution time. In contrast, the training times of the RF and XGBoost models are shorter than the LSTM model, and the prediction times are shorter than the KNN model. Meantime, the fusion model needs to perform feature fusion on the outputs of the four models, thus contributing to long training and prediction execution times. (Machine configuration, CPU: Intel Core i7-7700k 3.60 GHz; memory: 16.0 GB).

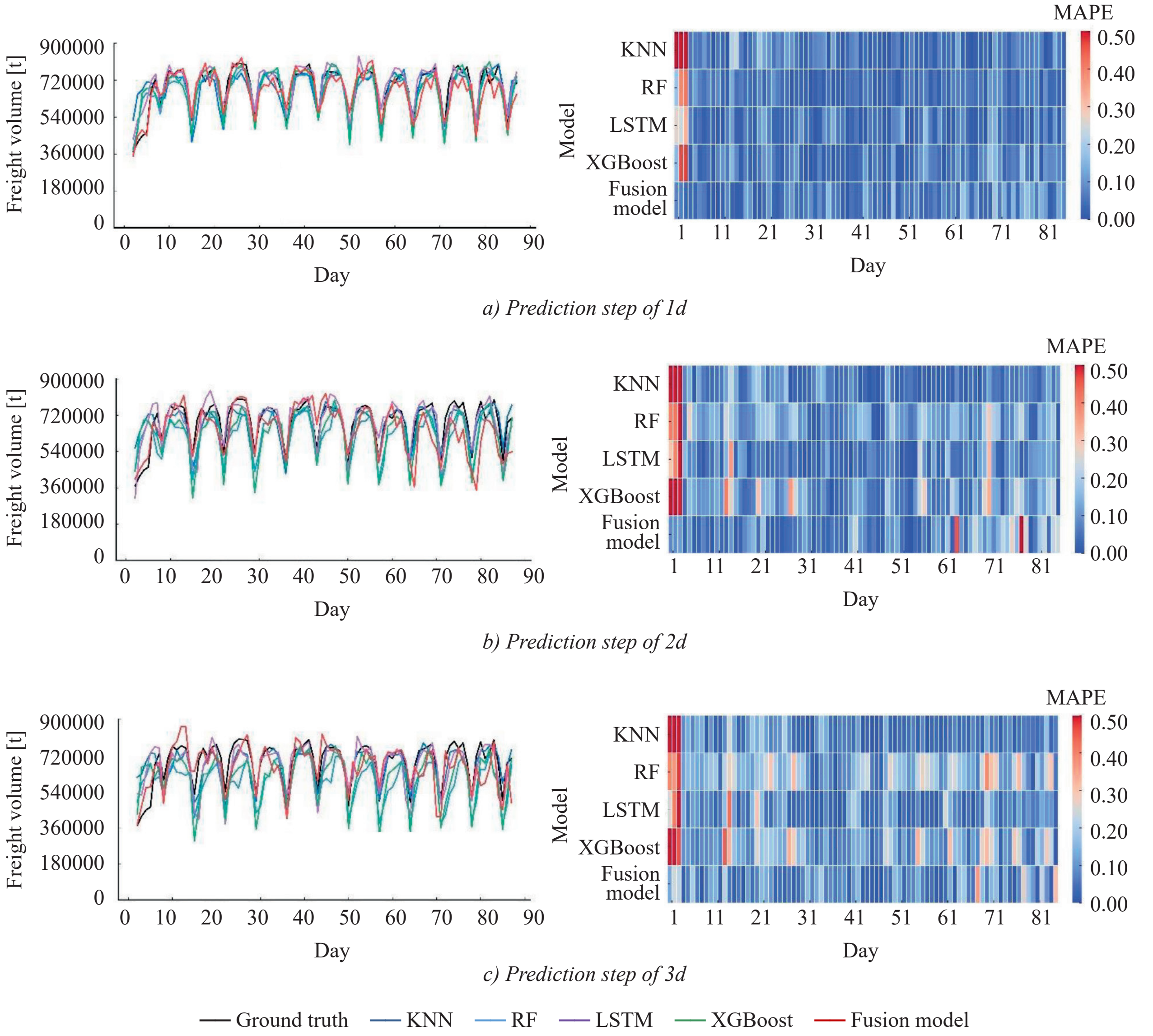


a) Prediction step of 1d

b) Prediction step of 2d

c) Prediction step of 3d

Figure 10 – Prediction results and absolute percent error heatmap for Nanjing



a) Prediction step of 1d

b) Prediction step of 2d

c) Prediction step of 3d

Figure 11 – Prediction results and absolute percent error heatmap for Suzhou

Table 10 – Training time of each model

Model	Training time [s]	Prediction time [s]
KNN	0.3823	0.9921
RF	2.4424	0.0232
LSTM	56.1782	0.4229
XGBoost	1.3229	0.1122
Fusion	60.7127	1.5491

5. CONCLUSIONS

In this study, the KNN model, RF model, LSTM model, XGBoost model and fusion model based on ridge regression are used to predict the freight volume of the Nanjing and Suzhou expressways with the introduction of meteorological information. The main contributions and findings can be summarised as follows.

We find that the prediction errors of the five models generally increased with increasing pre-diction step size, but the median and upper quartile of the MAPE increased slightly, and the maximum upper quartile was lower than 25%, which indicates that all models are relatively stable.

For different step sizes, in the forecast of freight volume in Nanjing, the prediction effects of the fusion, RF and XGBoost models are better than LSTM and KNN models. Among them, the LSTM model has the worst performance, which may be because there are two unbalanced sample data segments on the Nanjing freight data set, and the LSTN is prone to fall into local optimum, so it is more significantly affected than other models. When the prediction step is 1 day, the prediction accuracy of the fusion model reaches 96%, which is 3.2%, 2.9%, 6.5% and 8.7% higher than the other four models, respectively. In the forecast of freight volume in Suzhou, the fusion and LSTM models have better prediction performance. When the prediction step is 1 day, the accuracy reaches 93% and 90%, respectively.

The results demonstrate that the fusion model based on ridge regression can automatically adjust the weights, integrate the advantages of each model and overcome the limitations of a single prediction method. Accordingly, it presents the best prediction effect.

The fusion prediction method of expressway freight volume proposed in this study can provide a basis for the transportation management department to master the freight situation of expressways and the dynamic monitoring of freight volume. Furthermore, it also helps for exploring the degree of economic development activity in urban areas. In a follow-up study, we will further investigate the applicability of the prediction model at a smaller time accuracy, and more variables related to freight volume should be introduced to improve the accuracy of the model.

REFERENCES

[1] Cheng B, et al. Evolutionary game simulation on government incentive strategies of prefabricated construction: A system dynamics approach. *Complexity*. 2020;2020. DOI: 10.1155/2020/8861146.

[2] Yin S, et al. A data-driven fuzzy information granulation approach for freight volume forecasting. *IEEE Transactions on Industrial Electronics*. 2016;64(2):1447-1456. DOI: 10.1109/TIE.2016.2613974.

[3] Wu Y, Wang S, Zhang Y, Zhang J. A BP neural network model for the demand forecasting of road freight transportation system. *2021 International Symposium on Computer Technology and Information Science (ISCTIS)*. 2021. p. 264-268. DOI: 10.1109/ISCTIS51085.2021.00061.

[4] Yang F, et al. Forecast of freight volume in Xi'an based on gray GM (1,1) model and Markov forecasting model. *Journal of Mathematics* 2021;1:1-6. DOI: 10.1155/2021/6686786.

[5] Wang Y, Chen X, Han Y, Guo S. Forecast of passenger and freight traffic volume based on elasticity coefficient method and grey model. *Procedia - Social and Behavioral Sciences*. 2013;96:136-147. DOI: 10.1016/j.sbspro.2013.08.019.

[6] Zhang X, Wang S, Zhao Y. Application of support vector machine and least squares vector machine to freight volume forecast. *2011 International Conference on Remote Sensing, Environment and Transportation Engineering*. 2011. p. 104-107. DOI: 10.1109/RSETE.2011.5964227.

[7] Rashidi TH, Azevedo CML. Papers presented at the Transportation Research Board (TRB) 95th TRB Annual Meeting, Washington DC January 10-14, 2016. *Transportation*. 2016;43(6):951-953. DOI: 10.1007/s11116-016-9743-1.

[8] Smith BL, et al. An investigation into the impact of rainfall on freeway traffic flow. *83rd Annual Meeting of the Transportation Research Board, Washington DC, 2004*. CiteSeerX. 2004. DOI: 10.31224/osf.io/9xnzc.

[9] Chung E, et al. Does weather affect highway capacity. *Proceedings of the 5th International Symposium on Highway Capacity and Quality of Service*. 2006. p. 139-146.

[10] Ahmed MM, Ghasemzadeh A. The impacts of heavy rain on speed and headway behaviors: An investigation using the SHRP2 naturalistic driving study data. *Transportation Research Part C: Emerging Technologies*. 2018;91:371-384. DOI: 10.1016/j.trc.2018.04.012.

[11] Jia Y, Wu J, Xu M. Traffic flow prediction with rainfall impact using a deep learning method. *Journal of Advanced Transportation*. 2017;722:1-10. DOI: 10.1155/2017/6575947.

[12] Zou Y, Zhang Y, Cheng K. Exploring the impact of climate and extreme weather on fatal traffic accidents. *Sustainability*. 2021;13(1):390. DOI: 10.3390/su13010390.

[13] Singhal A, Kamga C, Yazici A. Impact of weather on urban transit ridership. *Transport Res a-Pol*. 2014;69:379-391. DOI: 10.1016/j.tra.2014.09.008.

[14] Zhan X, Hasan S, Ukkusuri SV, Kamga C. Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C: Emerging Technologies*. 2013;33:37-49. DOI: 10.1016/j.trc.2013.04.001.

[15] Zhan X, Zheng Y, Yi X, Ukkusuri SV. Citywide traffic volume estimation using trajectory data. *IEEE Transactions on Knowledge and Data Engineering*. 2016;29(2):272-285. DOI: 10.1109/TKDE.2016.2621104.

[16] Wu Y, Wang S, Zhang Y, Zhang J. A BP neural network model for the demand forecasting of road freight transportation system. *2021 International Symposium on Computer Technology and Information Science (ISCTIS)*. 2021. p. 264-268. DOI: 10.1109/ISCTIS51085.2021.00061.

[17] Chen C, Hu J, Meng Q, Zhang Y. Short-time traffic flow prediction with ARIMA-GARCH model. *2011 IEEE Intelligent Vehicles Symposium (IV)*. 2011. p. 607-612. DOI: 10.1109/IVS.2011.5940418.

[18] Abadi M. TensorFlow: Learning functions at scale. *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming*. 2016. DOI: 10.1145/3022670.2976746.

[19] Li H-J, Zhang Y-Z, Zhu C-F. Forecasting of railway passenger flow based on Grey Model and monthly proportional coefficient. *2012 IEEE Symposium on Robotics and Applications (ISRA)*. 2012. p. 23-26. DOI: 10.1109/ISRA.2012.6219110.

[20] Pace R K, et al. Spatiotemporal Autoregressive Models of Neighborhood Effects. *The Journal of Real Estate Finance and Economics.* 1998;17(1):15-33.

[21] Kamarianakis Y, Vouton V. Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches. *Transportation Research Record Journal of the Transportation Research Board*. 2003;1857(1):74-84. DOI: 10.3141/1857-09.

[22] Wang Y, Ma J, Zhang J. Metro passenger flow forecast with a novel Markov-Grey model. *Periodica Polytechnica Transportation Engineering*. 2020;48(1):70-75. DOI: 10.3311/PPtr.11131.

[23] Huang J, et al. Short-term travel time prediction on urban road networks using massive ERI data. *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. 2019. p. 582-588. DOI: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00138.

[24] Cai P, et al. A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. *Transportation Research Part C: Emerging Technologies*. 2016;62:21-34. DOI: 10.1016/j.trc.2015.11.002.

[25] Hong H, et al. Short-term traffic flow forecasting: Multi-metric KNN with related station discovery. *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. 2015. p. 1670-1675. DOI: 10.1109/FSKD.2015.7382196.

[26] Jiwon M, Kim D-K, Kho S-Y, Park C-H. Travel time prediction using k nearest neighbor method with combined data from vehicle detector system and automatic toll collection system. *Transportation Research Record*. 2011;2256(1):51-59. DOI: 10.3141/2256-07.

[27] Wu S, Yang ZZ, Zhu X, Yu B. Improved k-nn for short-term traffic forecasting using temporal and spatial information. *Journal of Transportation Engineering*. 2014;140(7):04014026. DOI: 10.1061/(ASCE)TE.1943-5436.0000672.

[28] Filmon, Mecit C. Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transportation Research Part C: Emerging Technologies*. 2016;66:61-78. DOI: 10.1016/j.trc.2015.08.017.

[29] Feng X, et al. Adaptive multi-kernel SVM with spatial–temporal correlation for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*. 2018;20(6):2001-2013. DOI: 10.1109/TITS.2018.2854913.

[30] Lu X, Gao J. Forecast of China railway freight volume by random forest regression model. 2015 International Conference on Logistics, Informatics and Service Sciences (LISS). 2015. DOI: 10.1109/LISS.2015.7369654.

[31] Chikaraishi M, et al. On the possibility of short-term traffic prediction during disaster with machine learning approaches: An exploratory analysis. *Transport Policy*. 2020;98:91-104. DOI: 10.1016/j.tranpol.2020.05.023.

[32] Zhou T, et al. An attention-based deep learning model for citywide traffic flow forecasting. *International Journal of Digital Earth*. 2022;15(1):323-344. DOI: 10.1080/17538947.2022.2028912.

[33] Tian Y, et al. LSTM-based traffic flow prediction with missing data. *Neurocomputing*. 2018;318:297-305. DOI: 10.1016/j.neucom.2018.08.067.

[34] Shu-xu Z, Bao-hua Z. Traffic flow prediction of urban road network based on LSTM-RF model. *Journal of Measurement Science & Instrumentation*. 2020;11(2).

[35] Do LNN, et al. An effective spatial-temporal attention based neural network for traffic flow prediction. *Transportation Research Part C: Emerging Technologies*. 2019;108:12-28. DOI: 10.1016/j.trc.2019.09.008.

[36] Liu Z, et al. A hybrid short-term traffic flow forecasting method based on neural networks combined with k-nearest neighbor. *Promet – Traffic & Transportation*. 2018;30(4): 445-456. DOI: 10.7307/ptt.v30i4.2651.

[37] Rajalakshmi V. Hybrid time-series forecasting models for traffic flow prediction. *Promet – Traffic&Transportation*. 2022;34(4): 537-549. DOI: 10.7307/ptt.v34i4.3998.

[38] Qiao Y, Wang Y, Ma C, Yang J. Short-term traffic flow prediction based on 1DCNN-LSTM neural network structure. *Modern Physics Letters B*. 2020;(3):2150042. DOI: 10.1142/S0217984921500421.

[39] Vidas M, Tubić V, Ivanović I, Subotić M. One approach to quantifying rainfall impact on the traffic flow of a specific freeway segment. *Sustainability*. 2022;14(9):4985. DOI: 10.3390/su14094985.

[40] Soua R, Koesdwiady A, Karray F. Big-data-generated traffic flow prediction using deep learning and dempster-shafer theory. *2016 International Joint Conference on Neural Networks (IJCNN)*. 2016. DOI: 10.1109/IJCNN.2016.7727607.

[41] *Highway capacity manual*. Washington, DC: TRB National Research Council; 2000.

[42] Pizzuti S, Moretti F, Panzieri S, Annunziato M. Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling. *Neurocomputing*. 2015;167:3-7. DOI: 10.1016/j.neucom.2014.08.100.

[43] Araújo MB, New M. Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*. 2007;22(1):42-47. DOI: 10.1016/j.tree.2006.09.010.

[44] Zhang W, et al. Vehicle traffic delay prediction in ferry terminal based on Bayesian multiple models combination method. *Transportmetrica A: Transport Science*. 2017;13(5):467-490. DOI: 10.1080/23249935.2017.1294631.

高宁，洪源伯，陈军飞，庞崇浩

引入气象信息的区域高速公路货运量预测算法

摘 要：
在后疫情时代，对公路货运量进行动态监测是一项重要的工作。高速公路在公路货运中占据重要的地位。为准确预测整个城市高速公路每日货运量，引入气象等信息，基于高速公路收费数据集，采用随机森林模型（RF）、极端梯度提升模型（XGBoost）、长短时记忆神经网络模型（LSTM）、K最近邻模型（KNN）等4种常用算法对货运量进行预测，并采用岭回归方法对各算法进行融合，融合模型可综合各预测算法的优点，具有更高的精度和鲁棒性。以2021年江苏省南京市和苏州市数据为例，利用过去一周的气象数据和货运量数据，分别对未来1天、2天、3天的货运量进行预测，并从预测精度和训练时长等方面对各算法的性能进行了对比分析，结果表明：南京市货运量预测中，随机森林算法和XGBoost算法的整体预测效果更好；苏州市货运量预测中，LSTM具有较高的精度；基于岭回归的融合预测方法综合了各预测算法的优点，在两个城市的货运量预测结果中，均表现出最优的效果。

关键词：
公路运输；货运量预测；机器学习；高速公路；气象信息