

A study on traffic congestion prediction based on random forest model

Shuo Sun^{1, #}, Hao Yan^{2, #}, Zejun Lang^{3, *, #}

¹ School of Mathematics and Statistics, Xinyang Normal University, Xinyang, China, 464000

² School of Environment, Nanjing Normal University, Nanjing, China, 210023

³ School of Mathematics, Hohai University, Nanjing, China, 211100

* Corresponding author: 13132708215@163.com

#These authors contributed equally.

Abstract. Enhancing the precision and minimizing false alarms in predicting urban traffic congestion is imperative for maintaining a smooth traffic flow in cities. This paper explores the patterns of road congestion from both a spatial perspective and a temporal perspective. In the spatial perspective, this paper finds that there are differences in the traffic congestion values of different roads through the box plots of the traffic congestion values of each road, and then finds that there is spatial correlation between the traffic congestion values of different roads through the Pearson correlation analysis. In terms of time, this paper finds that the congestion data is cyclical. Thus, it is necessary to establish a model based on both temporal and spatial characteristics. In this paper, the Lasso model, Ridge model and random forest regression model are used to fit the traffic congestion data, and the model is evaluated by using three evaluation indexes, namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), and the results of the study show that the random forest regression model is the most accurate in predicting the traffic congestion values.

Keywords: Traffic Congestion, Spatio-temporal Pattern, Pearson Correlation, Random Forest Regression.

1. Introduction

In the era of big data, the problem of urban traffic congestion is becoming increasingly serious with the acceleration of urbanisation. The impact of traffic congestion on people's travelling efficiency as well as the environmental impact of tailpipe emissions from increased vehicles need to be urgently addressed. Therefore, intelligent transport systems are beginning to receive more and more attention [1].

In recent years, a large number of scholars have conducted a lot of research on urban traffic congestion prediction. Chenhui Wang [2] The application of the combined model including ELM and EMD-ASPSO-GRU based and SARIMA-GA-Elman based, makes it possible to better fit the complex rule of change of traffic flow in traffic flow prediction; Sun Xianghai [3] et al. According to the short-term traffic flow on urban roads has non-linear characteristics, by analysing the correlation characteristics between traffic flow parameters and the spatio-temporal characteristics of multivariate time series. In addition, the use of two-system self-motivated threshold autoregressive model can be classified according to the different states of the traffic flow prediction, through these nonlinear time series model better reveal the intrinsic laws of the traffic flow; Yu Tao [4] Combining SVM and neural network can optimise the prediction algorithm. The combination of SVM and neural network can optimise the prediction algorithm. The combination algorithm is optimised by the optimal weighting rule to further improve the prediction accuracy. The theory of intelligence plays an important role in short-term traffic flow prediction; Li-Qiong Ku [5] et al. based on historical data, fusion speed, traffic flow in the region to construct the degree of road congestion indicators to build MM-SVR model, and use SVR and Adaboosting model for comparison experiments to prove the excellent performance of SVM model in traffic flow prediction.

However, most of the above studies are based on the time series prediction of traffic flow in a single cross-section, which only focuses on the changing law of traffic flow in time and lacks the discussion of the influence of spatial factors on traffic flow [6]. Therefore, this paper predicts traffic congestion based on a regression model and considering the dual factors of temporal pattern and spatial correlation of different roads. The simulation results optimise the accuracy and false alarm rate of congestion prediction and show the superiority of the prediction method.

2. Analysis of spatial and temporal patterns of traffic congestion data

Data from <https://www.chicago.gov>. <https://www.chicago.gov>.The data contains time, east-west midpoint coordinates of the road, north-south midpoint coordinates of the road, direction of the road, and traffic congestion values. Some of the data are shown in Table 1 below.

Table 1. Data sources

time	x	y	direction	congestion
1991-04-01	0	0	EB	70
1991-04-01	0	1	EB	18
1991-04-01	0	2	EB	31
1991-04-01	0	3	EB	18
...	...	...	...	...
1991-04-02	0	0	EB	40
1991-04-02	0	0	NB	47
...	...	...	...	...
1991-09-30	0	0	EB	48
1991-09-30	0	0	NB	40

2.1. Distribution of Traffic Congestion Data

Normal Distribution (Normal Distribution), also known as Gaussian Distribution; the probability density function of the normal distribution is bell-shaped, low at both ends, high in the middle, symmetrical, and its shape is determined by the mean and standard deviation. The histogram of the normality test of traffic congestion data is shown in Figure 1.

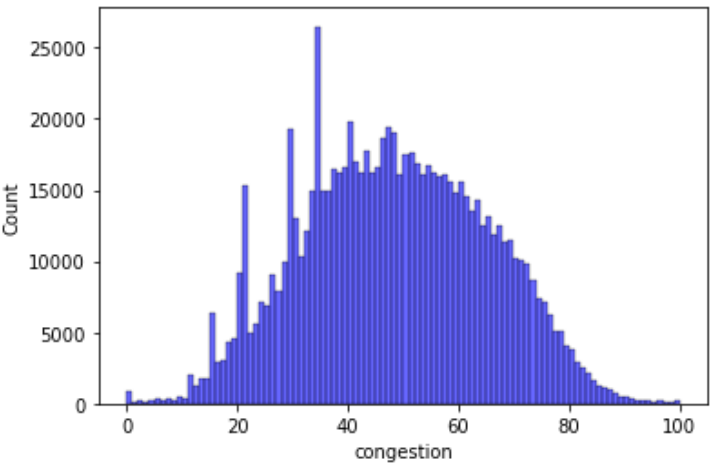


Figure 1. Histogram of normality test for traffic congestion data

Top Figure 1. shows a histogram of the normality test of the traffic congestion data, which basically shows a bell shape (high in the middle and low at the ends), the data is not absolutely normal but basically acceptable as a normal distribution. However, some of the traffic congestion values have much higher frequencies than other neighbouring values and are outliers.

First-order differencing, i.e., calculating the difference between each data point and the previous data point. Therefore, the first-order difference can be used to find out the outliers of certain traffic

congestion values which are much more frequent than other neighbouring values. The first order difference of the traffic congestion data is plotted as follows Figure 2 Shown.

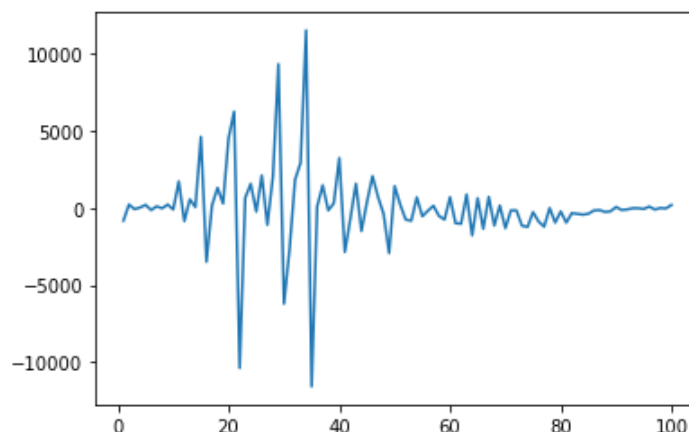


Figure 2. First order difference plot of traffic congestion data

Top Figure 2 shows the first order difference plot of the traffic congestion data and it is observed that the first order difference value fluctuates more at the points of 15, 20, 21, 29 and 34.

The traffic congestion anomalies identified by the first-order difference plot are marked with red bars, and the rest of the normal traffic congestion values are marked with yellow bars. Therefore, the normality test histogram of the traffic congestion data marked with anomalies is as follows Figure 3 shown.

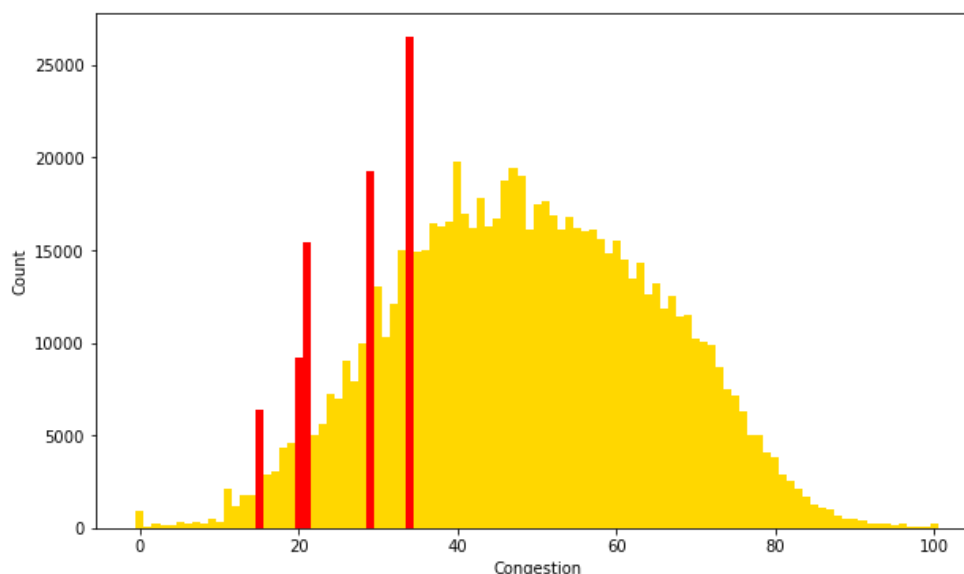


Figure 3. Histogram of normality test of traffic congestion data with marked outliers

Above figure 3 shows the histogram of the normality test of the traffic congestion data with marked outliers, it is observed that there are outliers and noise in the data, so simple linear regression fitting is no longer applicable, and this paper considers the application of the random forest algorithm which is insensitive to noise and has strong robustness.

2.2. Analysis of spatial patterns of traffic congestion

In this paper, x is the east-west midpoint of the road, and its value is $[0,1,2]$, and y is the north-south midpoint of the road, and its value is $[0,1,2,3]$. Each road has a different driving direction, in this paper, we use East Boulevard (EB) to indicate driving in the east direction, North Boulevard (NB) to indicate driving in the north direction, South Boulevard (SB) to indicate driving in the south direction, West Boulevard to indicate driving in the west direction, and Northeast (NE) to indicate driving in the north direction. Northeast (NE) denotes travelling in a northeast direction, Northwest

(NW) denotes travelling in a northwest direction, Southeast (SE) denotes travelling in a southeast direction, and Southwest (SW) denotes travelling in a southwest direction. The combination of coordinates and directions from the data in this paper is plotted as follows Figure 4 shown.

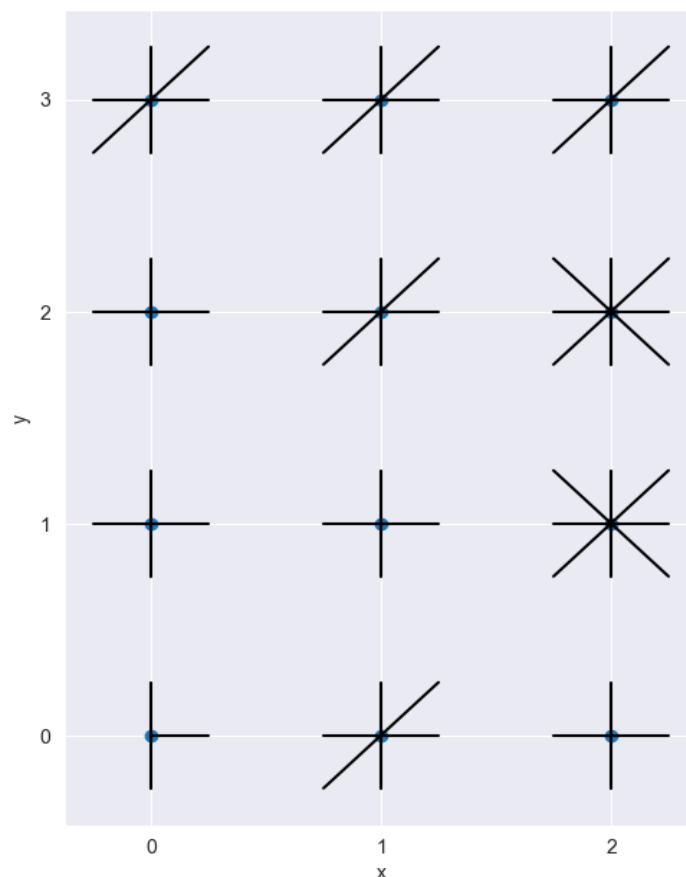


Figure 4. Distribution of road and direction combinations

Looking at the graph above, the x-axis has three coordinates 0, 1, and 2, and the y-axis has four coordinates 0, 1, 2, and 3, for a total of 12 coordinate positions, with up to eight directions on each ('EB' 'NB' 'SB' 'WB' 'NE' 'SW' 'NW' 'SE'), totalling 65 combinations of roads and directions. This means that on average a road has 5 or 6 directions.

Box-and-line diagram is a commonly used visualisation tool, which can show the scattering range and central position of multiple continuous data distributions, and is suitable for showing continuous variables that can take any value within a certain interval, and the width of the box in the box-and-line diagram reflects the fluctuation degree of the sample data to a certain extent. Therefore, this paper uses the box-and-line diagram to further analyse the differences between the traffic congestion values of different roads. In this paper, the traffic congestion values of 65 roads on 8 April are plotted as box plots, as shown in Figure 5 below.

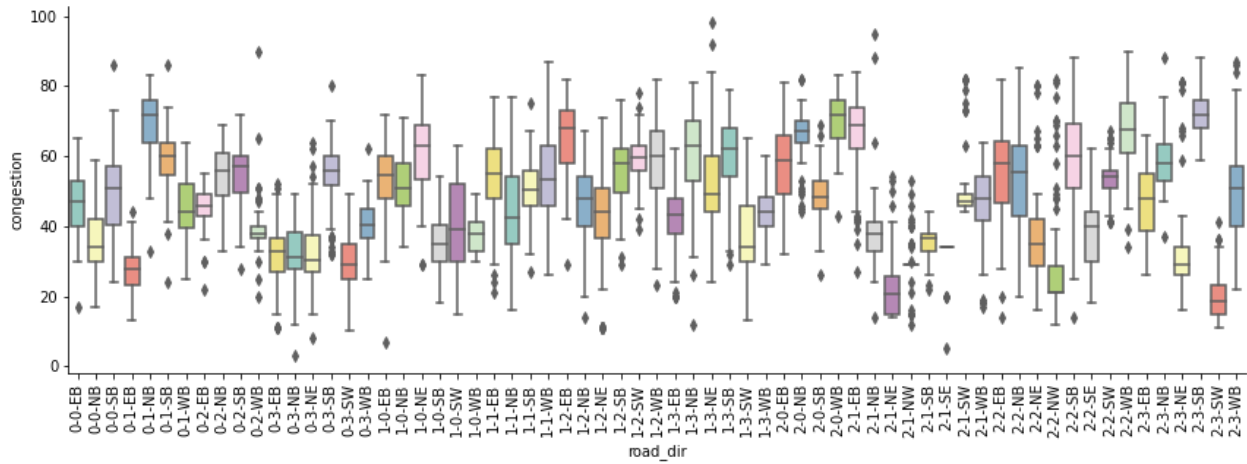


Figure 5. Box plot of traffic congestion values by road on 8 April

In this paper, the $(x=\alpha, y=\beta) + \gamma$ direction is used to denote the γ direction in $(x=\alpha, y=\beta)$ coordinates, for example, $(x=0, y=0)$ EB denotes the EB direction in $(x=0, y=0)$ coordinates. The width of the box is wider in $(x=1, y=0)$ SW, $(x=1, y=1)$ NB and $(x=2, y=2)$ NB, which shows a large dispersion of traffic congestion values. While $(x=0, y=2)$ EB, $(x=0, y=2)$ WB, $(x=2, y=1)$ NW, SB, SE, and SW boxes are shorter in width and have low dispersion. The median shows that $(x=0, y=1)$ NB, $(x=1, y=0)$ NE, $(x=1, y=2)$ EB, $(x=2, y=0)$ NB, WB, and $(x=2, y=1)$ EB have high average congestion values, while the average congestion values in the direction of $(x=0, y=1)$ EB, $(x=2, y=1)$ NE, and $(x=2, y=3)$ SW are low. The above analysis shows that there is a significant difference in congestion values between different roads.

Correlation coefficients can be used to measure the degree of linear correlation between two variables. The formula for calculating the correlation coefficient r of variables X and Y is shown in (1).

$$r_{X,Y} = \frac{CoV(X,Y)}{\sqrt{Var[X]Var[Y]}} \quad (1)$$

Where $Cov(X,Y)$ is the covariance of X and Y , $Var[X]$ and $Var[Y]$ are the variances of X and Y respectively. $-1 \leq r_{X,Y} \leq 1$ The covariance of X and Y is the covariance of X and Y respectively. When $r_{X,Y} > 0$ is used, it means that the two variables are positively and linearly correlated. When $r_{X,Y} < 0$ indicates that the two variables are negatively correlated, and when $r_{X,Y} = 0$ indicates that the two variables are not linearly correlated. It is usually considered that $|r_{X,Y}| \geq 0.8$ means that the two variables are strongly correlated, $0.5 < |r_{X,Y}| < 0.8$ means that the two variables are moderately correlated, $0.3 \leq |r_{X,Y}| < 0.5$ means that the two variables are weakly correlated, and $|r_{X,Y}| < 0.3$ means that the two variables are basically uncorrelated.

In order to understand the spatial correlation of different roads, this paper calculates the correlation coefficients of the traffic congestion values of different roads at 8:00 on 8 April and plots these correlation coefficients as a correlation coefficient matrix heat map as follows Figure 6 shows.

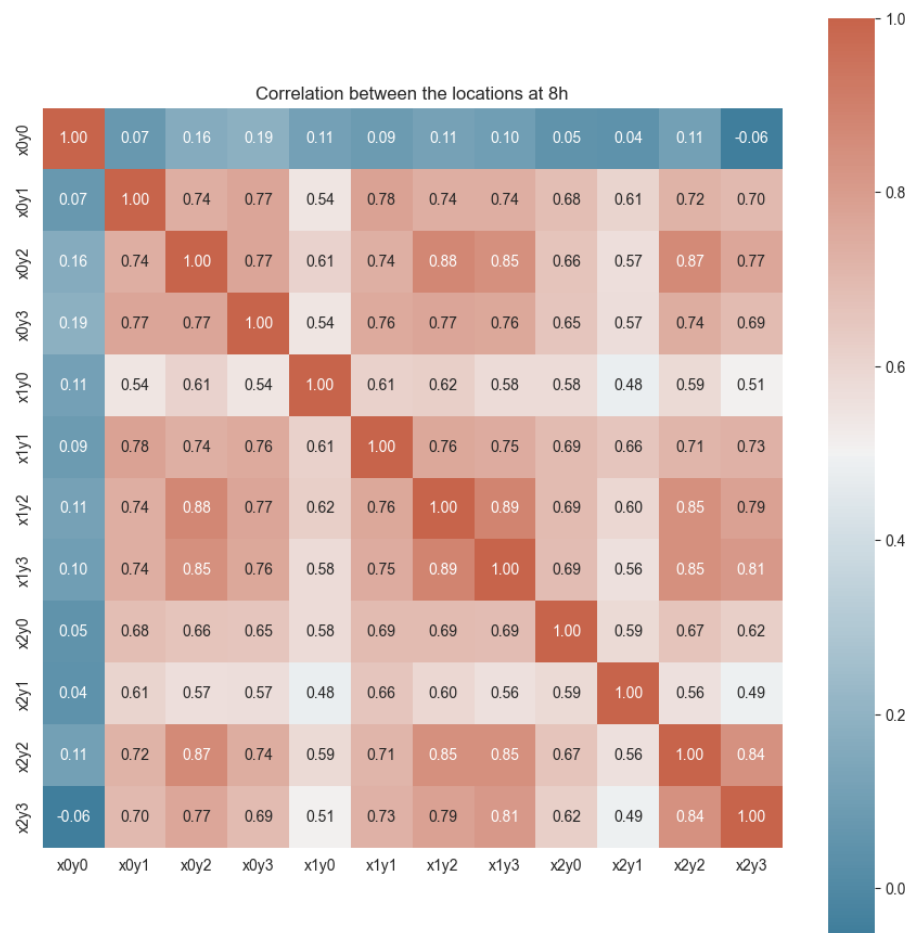


Figure 6. Heat map of correlation coefficient matrix of traffic congestion values for different roads

Figure 6 shows that the $(x=0,y=0)$ coordinate is essentially uncorrelated with the traffic congestion values of the other coordinates, while there is a strong correlation between the $(x=0, y=2)$, $(x=1, y=2)$, $(x=1, y=3)$ and $(x=2, y=2)$ coordinates. Similarly, there is a strong correlation between $(x=1, y=3)$ and the above coordinates. In contrast, the $(x=2, y=1)$ coordinate is weakly correlated with the $(x=1, y=0)$ and $(x=2, y=3)$ coordinates, and $(x=2, y=2)$ is strongly correlated with $(x=2, y=3)$. The remaining coordinates are moderately correlated.

In this paper, the time series decomposition function is used to decompose the traffic congestion time series into three parts: trend, seasonality and residual [7]. The trend part represents the long-term trend in the series, the seasonal part represents the periodic changes in the series, and the residual part represents the remaining part of the series. To use this function, you need to provide a time series data, and a period parameter, and the function will automatically analyze the time series data and return a decomposition result object, which contains the trend, seasonality and residual parts.

In this paper, the time series decomposition function is used to decompose the traffic congestion time series into three parts: trend, seasonality and residual. Among them, the trend part represents the long-term change trend in the series, the seasonality part represents the periodic change in the series, and the residual part represents the remaining part of the series. Firstly, the data is collected every 20 minutes, and due to the weekly seasonality in the time series, the number of time segments in a week with 20-minute interval data, i.e., 504 segments, will be taken as the periodicity parameter. Then, to reduce the trend and seasonality in the time series, the first weekly seasonal difference and the first non-seasonal difference in congestion levels are taken. This creates a new series where at each time interval t , the congestion level y becomes $y = (y_t - y_{t-504}) - (y_{t-1} - y_{t-504-1})$. Ultimately, the function automatically analyses the time series data and returns a decomposition result object with trend, seasonality and residual components, as shown in Figures 7-Figure 10 below.

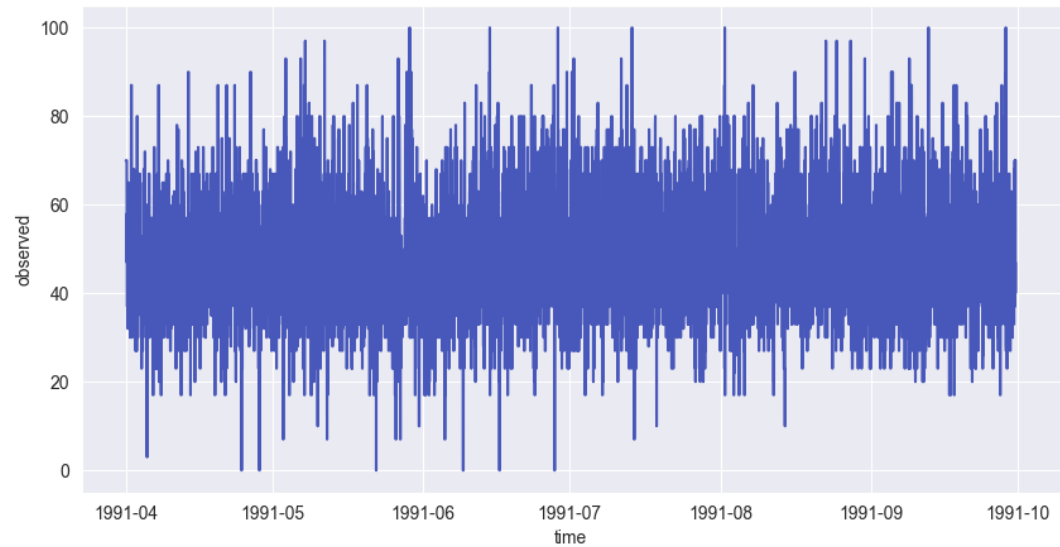


Figure 7. Plot of raw time series data

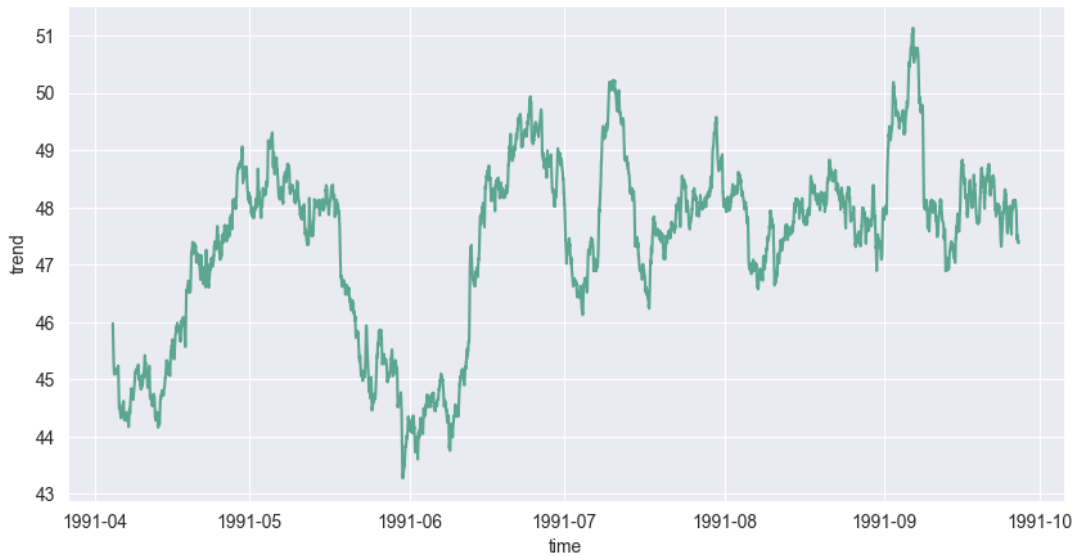


Figure 8. Decomposition of trend differences in congestion levels

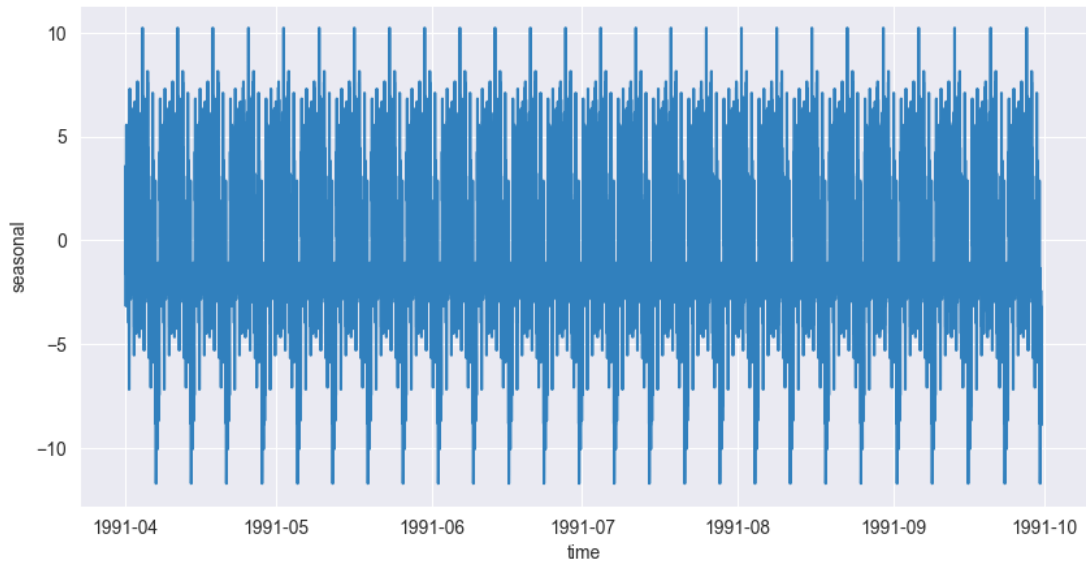


Figure 9. Decomposition of seasonal differences in congestion levels

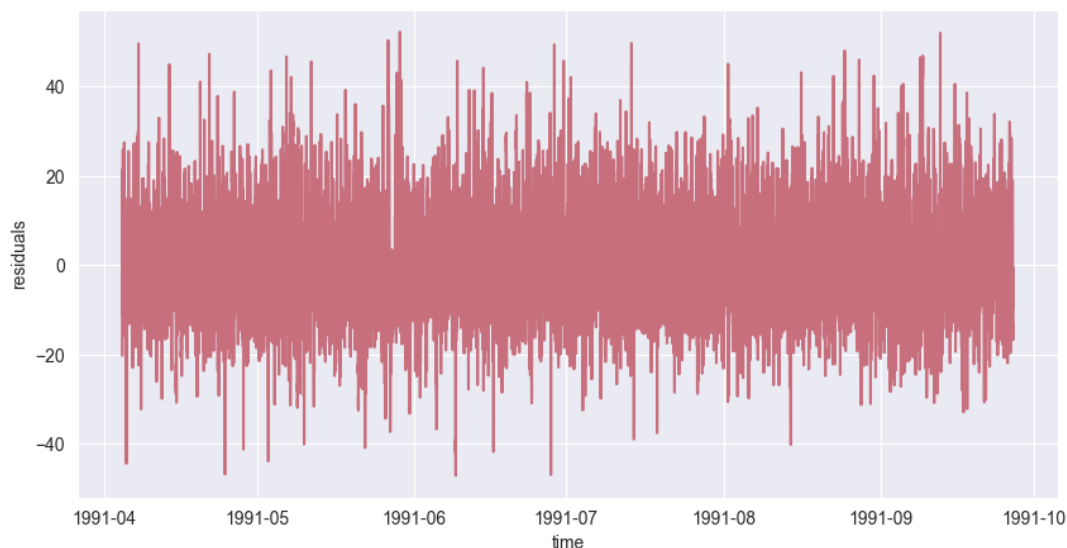


Figure 10. Residual decomposition plot for congestion levels

In Figure 9 of the Seasonal Variance Decomposition of Congestion Levels, it can be seen that most of the trend in the time series has been eliminated, with the mean stabilising and the data points very close to the x-axis. The range of variances in the seasonal graphs has also been significantly reduced. In addition, as a result of the above study, it was found that the congestion data shows a cyclical pattern and therefore the time factor needs to be taken into account when constructing the model.

3. Modelling of traffic congestion prediction considering spatio-temporal relationship

Random forest is an integrated learning method based on decision trees, which performs classification or regression by constructing multiple decision trees and outputting their patterns. Random forests have several key features [8]:

A. High Accuracy: Random Forests typically have extremely high accuracy among all current algorithms.

B. Large dataset processing power: Random forests can effectively run on large datasets, which makes it promising for large-scale data processing.

C. Handling high-dimensional features: Random Forest is able to handle input samples with high-dimensional features without the need for feature selection or dimensionality reduction. This means that it can handle datasets containing a large number of features without prior feature engineering.

D. Feature Importance Assessment: Random Forests are able to assess the importance of individual features for classification or regression problems, which is important for understanding data characteristics and improving models.

E. Robust and Noise Resistant: Since Random Forest constructs a decision tree by randomly selecting a subset of data and a subset of features, this makes it robust to cluttered datasets and noise.

F. Parallelised processing: the construction process of random forests can be easily parallelised, which helps to improve the efficiency of the algorithm's operation.

At the same time, according to the histogram of traffic congestion data normality test in Figure 3 above, it is found that there are outliers and noise in the data, so the simple linear regression fitting is no longer applicable, so this paper considers applying the Random Forest regression algorithm, which is insensitive to the noise and has a strong ability of robustness and anti-noise, to construct the prediction model of traffic congestion value.

3.1. model parameter

The model hyperparameters of the Random Forest regression-based model for predicting traffic congestion values developed in this paper are as follows Table 2 Shown.

Table 2. Random forest regression algorithm model hyperparameters

parametric noun	parameter value
n_estimators	10
criterion	Mse
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0
max_features	Auto
max_leaf_nodes	None
min_impurity_split	1e-07
bootstrap	True

3.2. Results of the model evaluation

MSE (Mean Square Error): The expected value of the squared difference between the predicted and actual values. This indicator gives higher weight to large errors, so the model pays more attention to large errors during the optimisation process. the smaller the MSE, the better the model's predictive ability. **RMSE (Root Mean Square Error):** The square root of the MSE, which is of more interest when the error is not very pronounced. the smaller the RMSE, the more accurate the model is. **MAE (Mean Absolute Error):** The average of the absolute errors. The MAE weights all individual differences equally on the mean compared to the MSE, so it highlights outliers better. MAE is more robust when there are outliers in the dataset. The smaller the value of MAE is taken, the more accurate the model is [9].

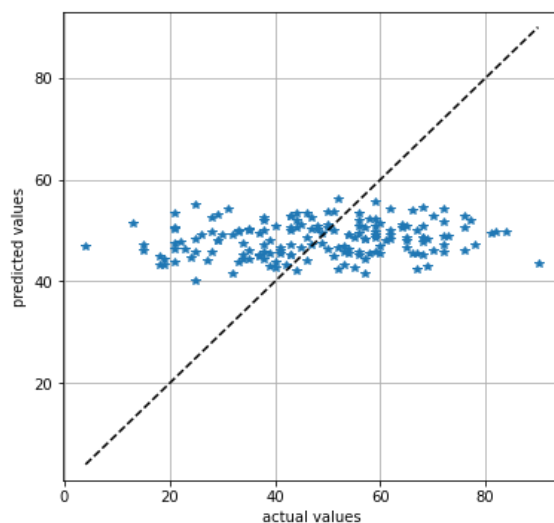
In order to find the model with the highest accuracy, this paper firstly used the Lasso model with lower computational complexity and the Ridge model for linear fitting, and the results showed that both models were not satisfactory. According to the result, this paper continues to use the random forest model for fitting, and the evaluation results of the three models are shown in the table below.

Table 3. Results of multiple regression model evaluation

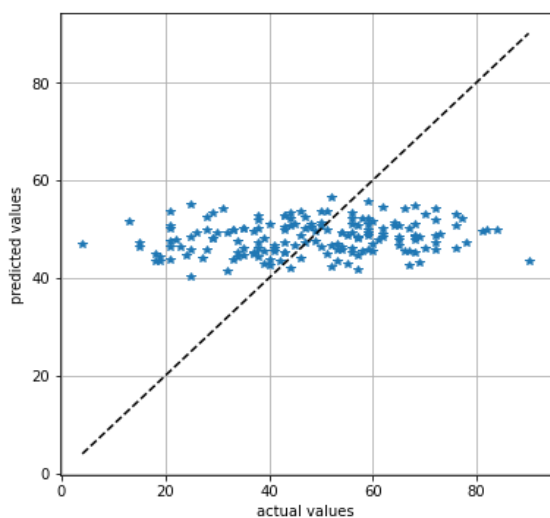
modelling	MAE	MSE	RMSE
Lasso	14.008	284.925	16.880
Ridge	14.008	284.925	16.880
Random Forest Regression	5.817	80.008	8.945

Observations Table 3, it is found that the three assessment indexes of random forest regression model are all minimum. Therefore, this paper finally chooses the random forest regression algorithm with the highest model accuracy to establish the traffic congestion prediction model.

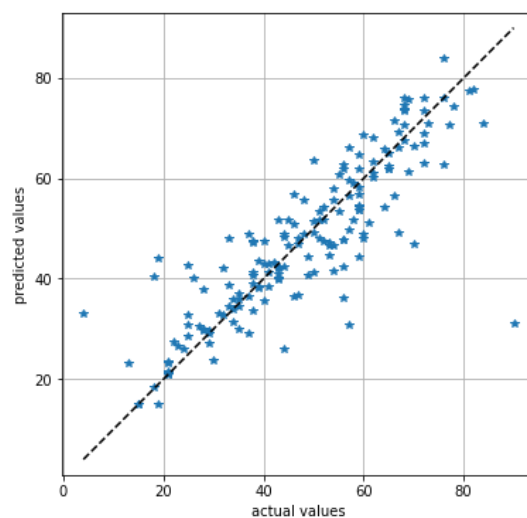
3.3. Model prediction results



(a) Ridge projections



(b) Lasso projections



(c) Random forest regression prediction results

Figure 11. Multiple regression model prediction results

Observe on Figure 11 Prediction results of various regression models , it is found that the predicted values of Lasso model and Rigde model are poorly correlated with the real values, and the prediction results are not ideal; on the contrary, the prediction results of Random Forest regression model show that the predicted values of traffic congestion are more strongly correlated with the real values, i.e., the prediction results are better. To sum up, this paper chooses the Random Forest Regression algorithm with more ideal model prediction results to establish the traffic congestion prediction model.

3.4. Feature Importance Ranking

Random forest algorithm is an integrated learning algorithm based on decision tree, which is used to train the decision tree by self help method resampling technique. Compared with machine learning algorithms such as neural networks, the random forest model has the advantages of strong interpretability, low data processing requirements, suitable for high-dimensional datasets, and fast training speed [10].

In this study, the traffic congestion value was used as the dependent variable, and a random forest model was built with a total of eight characteristic variables, namely time, coordinates, and direction,

respectively. The impurity that can be reduced when each observation is split in the tree is calculated, and after determining the importance of each influencing factor, it is sorted according to the order of the importance of the features from the largest to the smallest, and the results are as follows Figure 12 The results are shown below.

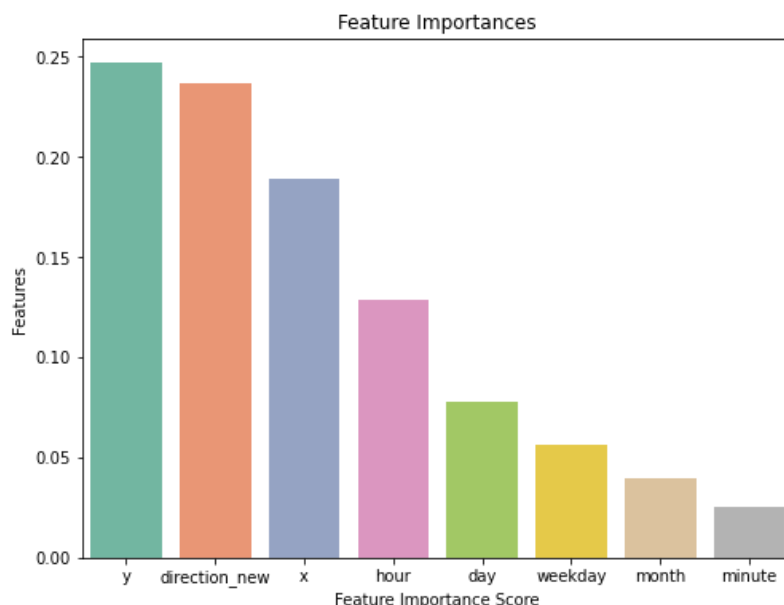


Figure 12. Percentage of feature importance for each feature variable

By means of the upper Figure 12, it is observed that the importance of the north-south midpoint coordinates of the road y, the east-west midpoint coordinates of the road x, and the direction features under the spatial features are significantly higher than the other temporal features, thus indicating that the spatial features are more important than the temporal features in the traffic congestion prediction in this paper.

4. Conclusion

This study commences with a meticulous examination of the spatial and temporal patterns inherent in traffic congestion data. In the spatial analysis, box plots are plotted to visualize the distribution of congestion values across various roads on April 8th. Notably, significant disparities in congestion levels are observed between different roads. To further investigate these spatial disparities, the study employs Pearson correlation to analyze the correlation between congestion values across roads. Turning to the temporal analysis, the study utilizes a time series decomposition function to dismantle the congestion time series into its constituent parts: trend, seasonality, and residuals. It is discovered that the congestion data exhibits a distinct cyclical pattern, highlighting the need to concurrently consider spatial and temporal factors in model construction. It is noteworthy that the traffic congestion data contains outliers and noise, rendering simple linear regression fitting inadequate. Consequently, this paper opts for the random forest regression algorithm, which exhibits robustness and insensitivity to noise. This algorithm is then utilized to predict traffic congestion values and is benchmarked against Lasso regression and Ridge regression. The results reveal that the random forest regression algorithm outperforms the others, achieving an MAE of 5.817, MSE of 80.008, and RMSE of 8.945, thereby establishing its superiority in prediction accuracy.

Forecasting traffic congestion requires a meticulous consideration of multiple factors that jointly influence the flow of vehicles. These include variables such as weather patterns, road conditions, the ebb and flow of traffic, the intricate structure of the road network, the availability and efficiency of public transport services, urban planning considerations, and land utilization practices. Given the intricate nature of these factors, it is crucial for pertinent departments to adopt tailored measures. This includes the establishment of intelligent transport systems that harness technological advancements

to streamline traffic flow, the reinforcement of traffic enforcement measures to ensure compliance with regulations, the optimization of public transport services to provide a viable alternative to private vehicles, the enhancement of road connections to improve connectivity and reduce bottlenecks, the implementation of public transport priority strategies to encourage its utilization, and the formulation of urban evacuation strategies to manage congestion during peak hours. By executing these strategies, the adverse impact of traffic congestion can be significantly mitigated, leading to a more efficient urban transportation system and an enhanced quality of life for residents.

References

- [1] REN Shanshan, BUAI Xinlei. An urban traffic congestion prediction method based on time series modelling [J]. Information and Computer (Theoretical Edition), 2023, 35 (10): 1 - 3.
- [2] Wang Chenhui. Research on short-time traffic flow prediction method based on combination model [D]. Lanzhou University of Science and Technology, 2021.
- [3] Sun Xianghai, Liu Tanqiu. Short-term traffic flow prediction of urban roads based on nonlinear time series model [J]. Journal of Civil Engineering, 2008, (01): 104 - 109.
- [4] Yu T. Short-time traffic flow prediction and implementation based on SVM and BP neural network [D]. Nanjing University of Posts and Telecommunications, 2018.
- [5] KU Li-Qiong, SONG Zukang, YANG Yang. Traffic congestion prediction based on improved support vector machine model [J]. Software Guide, 2019, 18 (12): 61 - 64+68.
- [6] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model [J]. Systems Engineering-Theory and Practice, 2010, 30 (1): 158 - 160.
- [7] Song Xudong, Ren Mengxue. Short-time traffic flow prediction based on combinatorial model [J] Computer Simulation, 2022, 39 (07): 156 - 160+406.
- [8] SUN, YAO Baozhen, BAI Zijian. Evaluation of traffic congestion diversion effect based on random forest model [J/OL]. Journal of Jilin University (Engineering Edition), 1 - 8 [2024-03-04].
- [9] Qiu Shishong, Lu Baichuan, Ma Qinglu, Zou Wei, Zhang Qin. Traffic flow prediction based on spatio-temporal characterization and data fusion [J]. Journal of Wuhan University of Technology (Information and Management Engineering Edition), 2015, 37 (02): 156 - 160+178.
- [10] BAI Xiaoman, FENG Yongxiang, LI Leixiao et al. Optimised random forest prediction model for urban road congestion [J]. Science, Technology and Engineering, 2021, 21 (26): 11205 - 11211.