A Study on Short-Term Traffic Flow Prediction Based on Random Forest Regression

Jiong Wu *

College Of Mechanical And Electrical Engineering, Hai Yue Academy, Harbin Engineering University, Harbin, China, 150001

* Corresponding Author Email: 15184129088@163.Com

Abstract. Accurate and rapid prediction of short-term traffic flow is of great significance for improving the efficiency of traffic management as well as alleviating the pressure of urban traffic. In this paper, the traffic flow data are first analysed in depth in terms of spatial and temporal patterns. In the time dimension, through the daily traffic flow statistics line graphs of the four junctions, it is found that the traffic flow of the No.1 junction shows an obvious trend of increasing year by year, while the traffic flow of the No.2, No.3, and No.4 junctions grows more gently without any significant incremental pattern. In addition, the weekly traffic flow line graphs reveal a clear pattern: the traffic flow from Monday to Friday is relatively high, while the traffic flow on Saturdays and Sundays is relatively low. In the spatial dimension, the violin distribution plots of the traffic flow in each junction reveal that the traffic flow in each junction shows a skewed distribution. Therefore, this paper employs Quantile Transformer for data preprocessing to eliminate this skewness. Next, this paper explores the correlation of traffic flow between different junctions using Pearson's correlation coefficient, and the results show that the traffic flow between junctions is extremely correlated. In order to predict the traffic flow, this paper tries three models, Lasso, Ridge and Random Forest Regressor, respectively, and searches for the optimal hyperparameters of the above models through cross validation grid search. Finally, four evaluation metrics, namely, Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and Coefficient of Determination R2, were used to comprehensively evaluate the prediction accuracy of the models. Eventually, it returned that the Random Forest Regressor model performed the best in terms of prediction results, with an MAE of 17.315, an MSE of 4.161, an RMSE of 2.458, and an R2 of 0.96.

Keywords: Short-term Traffic Flow, Spatio-temporal Analysis, Data Destratification, Pearson Correlation Analysis, Random Forest Regressor.

1. Introduction

With the acceleration of urbanisation and the growth of public travel demand, the traditional urban transport system is facing increasingly serious problems such as huge number of equipment, invisible waste of energy, lagging passenger flow monitoring, single fault information, and dispersed information municipalities. In such a background, intelligent transport system came into being, and gradually become an important means to alleviate urban traffic pressure and improve traffic management efficiency [1]. Among them, the diagnosis and prediction of the current urban traffic congestion is an important part of the intelligent transport system [2].

Short-term traffic flow prediction, i.e., the use of historical and real-time traffic data to predict changes in traffic flow or traffic state in a short period of time in the future (usually a few minutes to a few hours). Currently there are more model methods used for short-term traffic flow prediction, mainly including: nonlinear models [3], deep learning models [4], traditional traffic assignment models [5] and simulation models [6]. Linear models have strong extraction ability for trend and periodic features in historical traffic flow data, but have poor extraction ability for disturbance term features such as short-term fluctuations in the data, so they are more suitable for medium- and long-term prediction. Chaos model uses a large number of parameters to describe the traffic flow, which has a large amount of computation and limited accuracy, and the feature matching model (e.g. K-nearest neighbour algorithm) mainly achieves the purpose of extracting data features by grouping the

historical data, which is mainly aimed at the more obvious cyclical and trending features, and is not applicable to the short-term traffic flow prediction [7].

Nonlinear model may be affected by local extreme points when dealing with complex traffic flow prediction problems, resulting in inaccurate prediction results. In addition, parameter selection and optimization of nonlinear models may require more domain knowledge and experience and may be more difficult for non-experts. As for the deep learning model: Deep learning model requires a large amount of data to train and adjust parameters, and the acquisition and annotation of traffic data is expensive. In addition, the black-box nature of deep learning models makes them less interpretable and difficult to understand the decision-making process inside the models, limiting their application in the field of transportation. As for the traditional traffic allocation model: The traditional traffic allocation model may oversimplify the complexity of traffic flow, and cannot capture the nonlinear and dynamic characteristics of traffic flow well, resulting in low prediction accuracy. In addition, the parameter setting and calibration of traditional traffic allocation models require a lot of experience and field research, which may not be suitable for complex traffic systems.

In summary, this paper presents a high-accuracy, low-complexity machine learning model for short-term traffic flow prediction. It uniquely analyzes the spatio-temporal pattern of traffic flow using Pearson's correlation coefficient, enhancing forecast accuracy. Additionally, it employs Quantile Transformer to address skew distribution, improving model stability and accuracy. The paper compares multiple regression models, identifying optimal hyperparameters through cross-validation, providing valuable insights for practical applications.

2. Analysis of spatial and temporal patterns of traffic flow data sources

Source of data: https://www.kaggle.com/.The data totalled 48,120 samples and the dataset was the hourly traffic flow for four different junctions, which contained ID serial number, date and time, Junction information, and vehicle flow (Vehicles). Some of the data is shown in Table 1 below.

Table 1. Data presentation.

ID	Date Time	Junction	Vehicles
0	2015-11-01 00:00:00	1	15
1	2015-11-01 01:00:00	1	13
2	2015-11-01 02:00:00	1	10
3	2015-11-01 03:00:00	1	7
...	...	...	...
48120	2017-6-30 22:00:00	4	22
48121	2017-6-30 23:00:00	4	12

2.1. Time pattern analysis

The time series of traffic flow for the four junctions are plotted [8], as shown in Fig. 1. The red folded line represents the traffic flow of junction 4, the green folded line represents the traffic flow of junction 3, the blue folded line represents the traffic flow of junction 1, and the yellow folded line represents the traffic flow of junction 2.

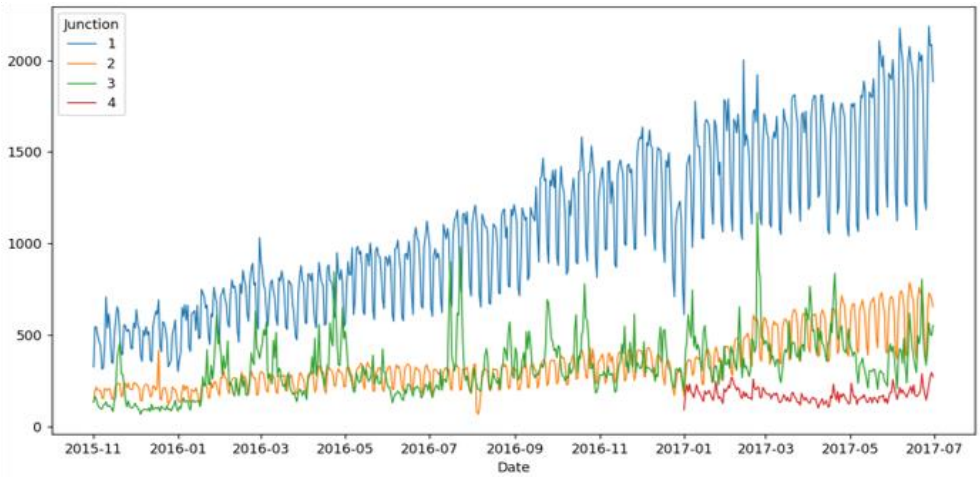


Figure 1. Line graph of daily traffic counts.

Looking at Figure 1 above, it is found that junction 4 only has data for 2017.Vehicle flows in junction 1 show an increase year on year; vehicle flows in junctions 2, 3 and 4 fluctuate during this period, with no significant incremental increase. 22 12 Now, the four-junction traffic flow by month and year, respectively, to draw the month traffic flow statistics graph and year traffic flow statistics graph. They are shown in Figures 2 and 3, respectively. The red bar represents No. 4 junction traffic, the green bar represents No. 3 junction traffic, the yellow bar represents No. 2 junction traffic, and the blue bar represents No. 1 junction traffic.

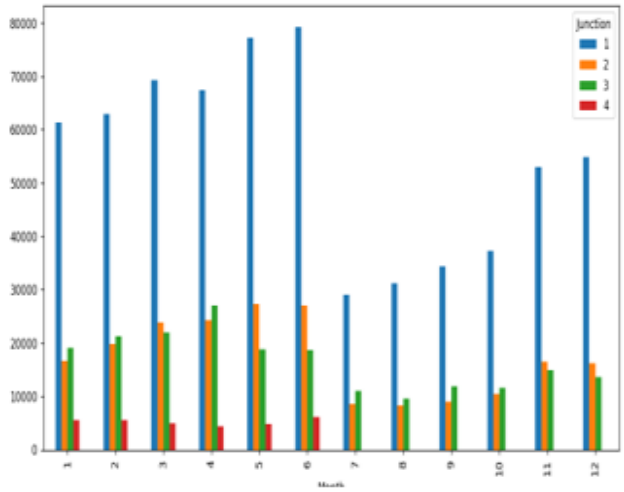


Figure 2. Traffic counts by junction month.

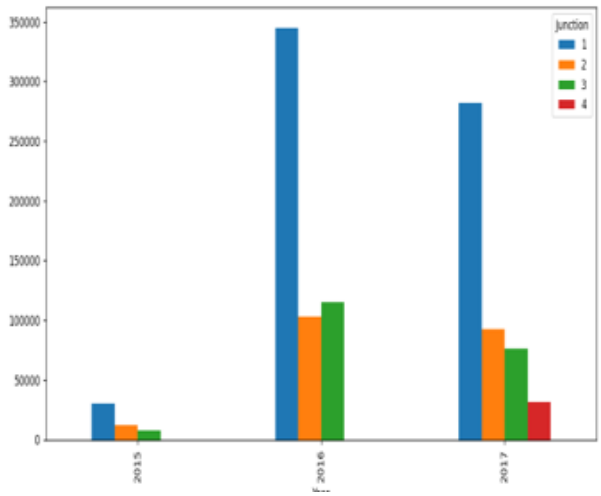


Figure 3. Traffic counts by junction year.

Looking at Figure 2, it is observed that the average traffic flow in the first six months is much larger than the second six months, especially in junction 1 and 2 where there is a large influx of traffic in the first six months. The reason for this is most likely due to the movement of people around the summer months. Looking at Figure 3, it is found that the traffic flow seems to be on an upward trend in all junctions except for junction 4 where only the traffic flow for the first half of the year is recorded. The four-junction traffic is now counted on a daily basis and the daily traffic statistics are plotted. As shown in Figure 4.

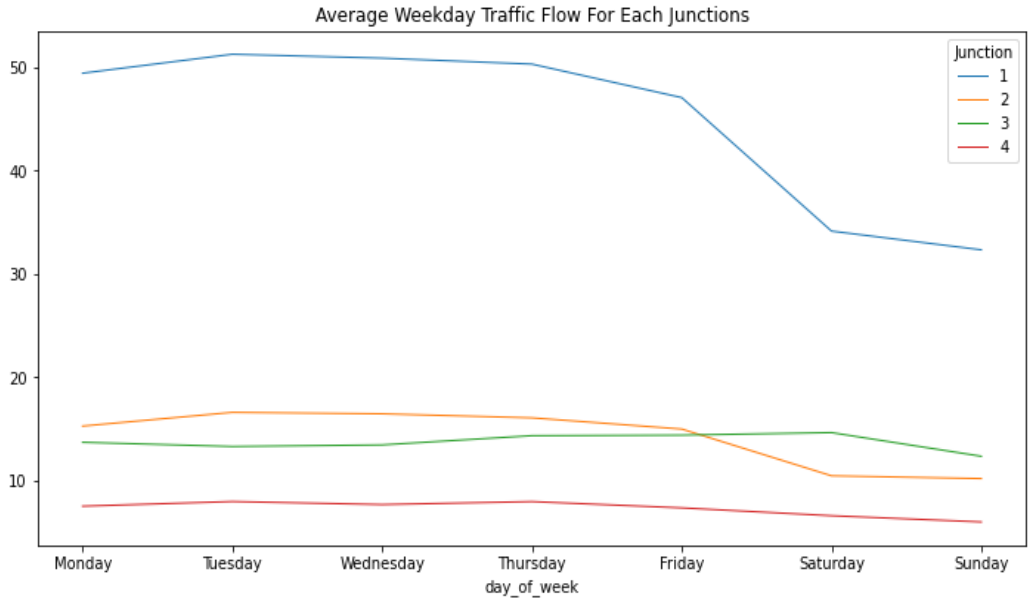


Figure 4. Weekly traffic flow line graph for each junction.

Observing Figure 4 above, it is found that the traffic flow is higher from Monday to Friday and lower on Saturday and Sunday. The reason for this is mainly due to people's travelling habits and daily living arrangements. For example, travel demand is higher on weekdays: Monday to Friday is the working time for most people. During this time, people need to travel to their workplaces, schools or other necessary activities, thus generating a large amount of commuter traffic. This regular travel demand results in relatively high traffic volumes on weekdays. Less demand for travelling on non-working days: Saturday and Sunday are days of rest for most people. During this time, people may choose to rest at home, travel or engage in other recreational activities. Although the total number of trips on weekends may not be low, the volume of traffic will be relatively small compared with that on weekdays due to the dispersed nature of travelling time and destinations.

2.2. Analysis of spatial patterns

Violin diagram is a type of chart used to visualise the distribution of numerical data, combining the advantages of box-and-line and kernel-density plots, and is usually used to compare the difference in distribution between multiple groups, or to show the distribution of a variable under different types. In order to analyse the distribution of traffic flow in each junction, the traffic flow in each junction is shown by violin diagram as shown in Fig. 5.

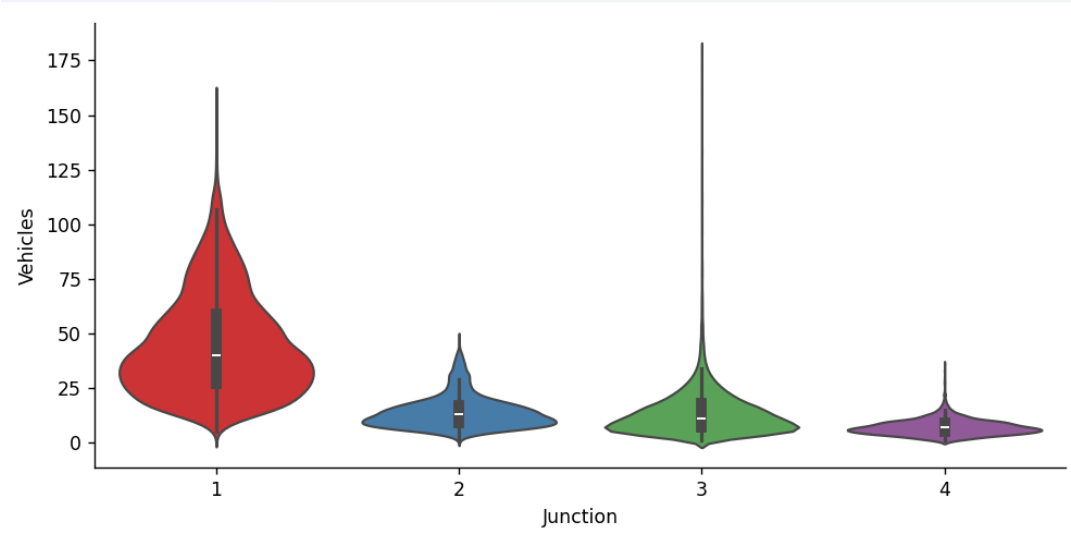


Figure 5. Distribution of violins in traffic flow by junction.

Observing the above Fig. 5, it is found that there are differences in the traffic flow in the four junctions, which indicates that the traffic flow prediction model should take the spatial information into consideration. Meanwhile, it is found that the traffic flow of the four junctions shows a skewed distribution [9], so the traffic flow of the four junctions should be processed separately to eliminate the skewness of the data. The Quantile Transformer method is a nonlinear transformation method that removes skewness from data by transforming it into a uniform distribution over a specified number of quartiles. The Quantile Transformer method smoothes the distribution of the anomalies by performing a rank transformation and is able to suffer less from outliers than scaling. Specifically, the values of the original data are mapped to new values that have the same cumulative distribution function values at specified quantile positions. In this way, the data is mapped to a uniform distribution from zero to one. In the following, the data are processed by Quantile Transformer method to eliminate the bias of the data for the above four junctions of traffic flow. The processed data is shown in Fig. 6, after Quantile Transformer processing, all the four junction traffic volumes eliminate the skewness and show normal distribution.

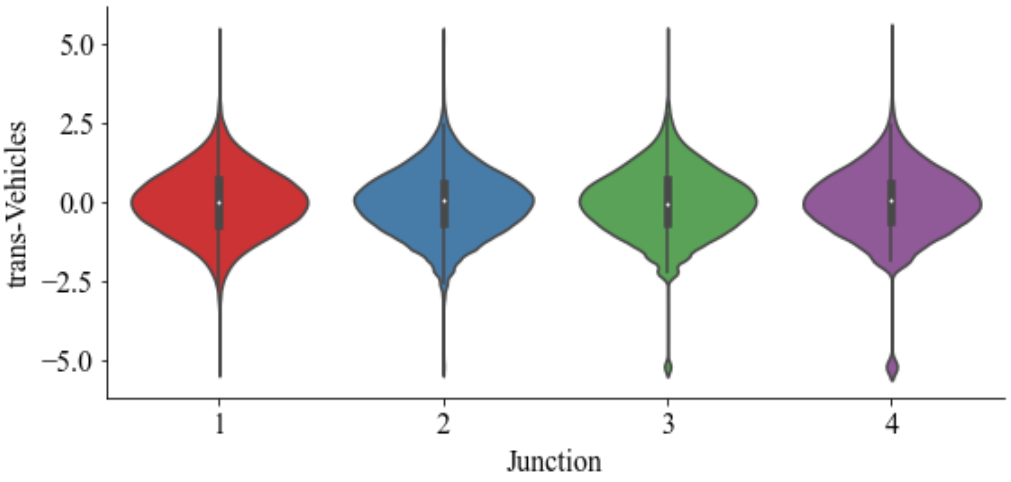


Figure 6. Violin distribution of the deflated state of traffic flow in each junction.

Correlation Coefficients (CC) can be used to measure the degree of linear correlation between two variables [10]. The formula for calculating the correlation coefficient r of variables X and Y is shown in (1).

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

Where is the covariance of X and Y, and, are the variances of X and Y respectively., at that time indicates that the two variables are positively correlated linearly, at that time indicates that the two variables are negatively correlated linearly, and at that time indicates that the two variables are not linearly correlated linearly. It is usually considered that, then the two variables are strongly correlated; then the two variables are moderately correlated; then the two variables are weakly correlated; then the two variables are basically uncorrelated. From the basic principle of traffic flow, the traffic flow between different junctions may be affected by each other. In order to analyze the correlation of traffic flow between different junctions, the following calculates the Pearson correlation coefficient of traffic flow between different junctions and plots the heat map of correlation coefficient. As shown in Figure 6, the Pearson correlation coefficient heat map between the daily 8 o'clock statistical traffic flow of the three junctions will be shown; as shown in Figure 7, the Pearson correlation coefficient heat map between the 2017 traffic flows of the four junctions will be shown.

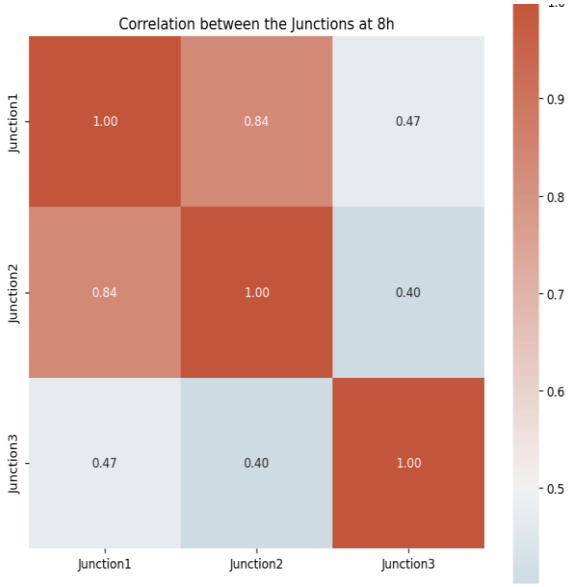


Figure 7. Heat map of Pearson's correlation coefficients between statistical traffic flows at 8:00 p.m. daily for junction123.

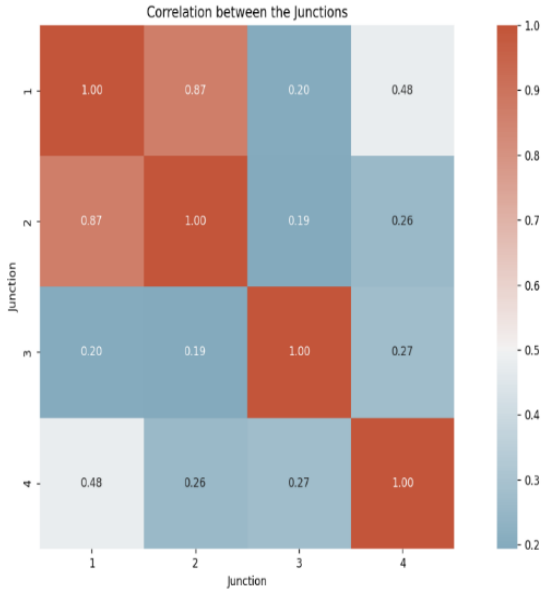


Figure 8. Heat map of Pearson's correlation coefficients between statistical traffic flows at 8:00 p.m. daily for junction4.

Observing the above Figures 7and 8, it is found that the traffic flow in junction one and junction two have strong correlation. That is, the traffic flow between each junction has spatial correlation, and the spatial information junction should be taken into account when establishing the traffic flow prediction model.

3. Modelling of Traffic Flow Prediction

3.1. Data segmentation

In this paper, a total of 48,120 samples are randomly divided into a training set and a test set in the ratio of 8:2, where the training set contains 38,496 samples and the test set contains 9,624 samples. The training set is used to train the model and the test set is used to evaluate the model performance. By splitting the dataset into training and test sets, it can be ensured that the model is evaluated on unseen data, thus avoiding model overfitting.

3.2. Cross-validation grid search

Hyperparameters are the parameters that need to be set before the machine learning model is trained, and the selection of these parameters has a significant impact on the performance of the model. In this paper, we choose the grid search method, i.e., we find the optimal combination of parameters by exhaustively searching a specified grid of parameter values. Cross-Validation performs multiple rounds of validation by dividing the dataset into multiple subsets, i.e., multiple estimates of model performance can be obtained. The average of these estimates provides a more accurate assessment of the model's true performance and evaluates its stability. Therefore, in this paper, we find the optimal hyperparameters of the model through the Grid Search CV function, in which 3-fold cross-validation is chosen to evaluate the performance of each combination of hyperparameters, i.e., the training set is divided into 5 copies, and each time, 4 copies of the training set are used for training, and the remaining 1 copy is used for validation. This process is repeated 5 times. The average of the 5 cross-validation model accuracies for a particular combination of hyperparameters is used to represent the accuracy of the model for that combination of hyperparameters; and mean_ squared_ error is specified as the scoring function to assess the model accuracy. In this paper, we have tried Lasso, Rigde, Random Forest Regressor, three regression prediction models respectively; The final optimal hyperparameters for the 3 models were: lasso___ alpha=0.0, ridge__ alpha=1.0, random forest regressor__ min_ samples _leaf=3.

3.3. Comparison of model results

Different evaluation indexes have different focuses, so this paper combines multiple indexes to evaluate the above models comprehensively to obtain more comprehensive model performance evaluation results. The values of the evaluation indexes for the three models are shown in Table 2 below

Table 2. The evaluation indexes for the three models.

Model	MSE	RMSE	MAE	R^2
Lasso	168.047	12.963	9.636	0.615
Ridge	168.047	12.963	9.636	0.615
Random Forest Regressor	17.315	4.161	2.458	0.960

3.4. Model hyperparameters.

Observing the values of the evaluation metrics for the x3 models in the above table, it is found that the Random Forest Regressor model performs the best, and the hyperparameters of this model are selected as shown in Table 3 below:

Table 3. Hyperparameters of the Random Forest Regressor model.

Parameter Name	Parameter Value
cross-validation	5
Guidelines for the evaluation of node	mes
Maximum proportion of features considered in the division	None
Minimum number of samples for internal node splits	2
Minimum number of samples for leaf nodes	1
Minimum weight of samples in leaf nodes	0

3.5. Presentation of model results

Scatter plots comparing true and predicted values visualize model performance and reveal errors. A tight scatter around the 45-degree line suggests accurate predictions, while a scattered plot indicates larger errors. These plots also aid in spotting outliers or specific error patterns, enabling further analysis and model optimization. Initial bias and variance assessments can be made by observing the plot's distribution and shape. Adjusting model parameters or trying different algorithms can reduce bias, while stabilizing the model's fit can minimize variance.

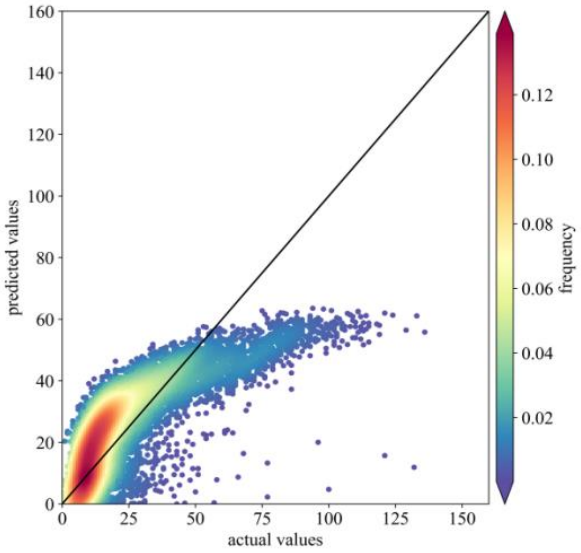


Figure 9. Scatter plot of density between true and predicted values of lasso model.

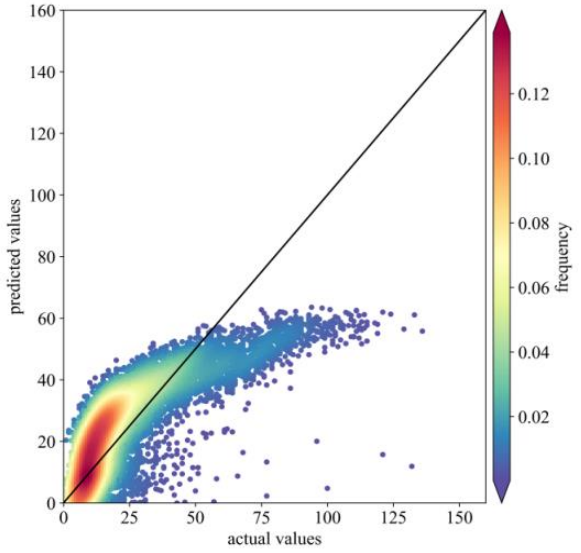


Figure 10. Scatter plot of density between true and predicted values of rigde model.

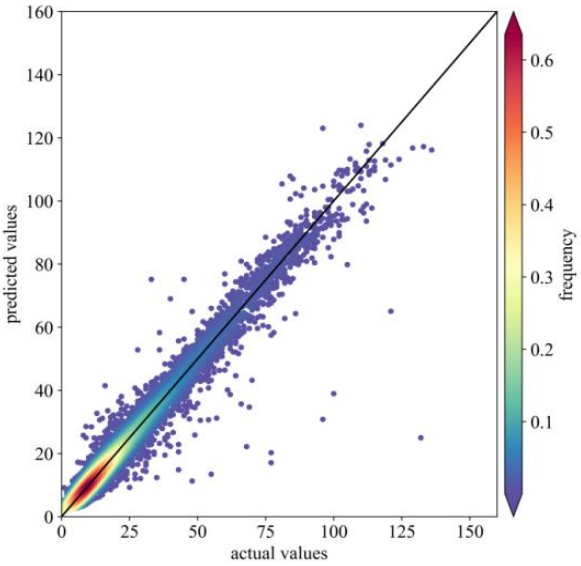


Figure 11. Scatter plot of density between true and predicted values of Random Forest Regressor model.

Observing the above Figures 9, 10, 11, it is found that the Random Forest Regressor model has its scatters closely surrounding the 45-degree line, which indicates that the Random Forest Regressor model has a better prediction performance. The Lasso, rigde model scatters are scattered or far away from the 45-degree line, which indicates that the model has a larger prediction error.

4. Conclusions

This study deeply explores traffic flow patterns. Over time, Junction No. 1's flow increases yearly, with flatter growth in others. Traffic peaks in the first half of each year, and is higher on weekdays. Spatial analysis reveals skewed traffic, corrected by Quantile Transformer. Traffic flow strongly correlates between junctions. Models tested include Lasso, Ridge, and Random Forest Regressor, with hyperparameters optimized through cross-validation. Random Forest Regressor best predicts traffic, with MAE of 17.315, MSE of 4.161, RMSE of 2.458, and R2 of 0.96.

In actual short-term traffic flow prediction, it is affected by many factors. The factors will change the fluctuation of traffic flow to a large extent. Therefore, the next step is to consider the introduction of a comprehensive range of factors affecting traffic flow, to further improve the accuracy of the model prediction, and to provide more scientific and reasonable decision support for the traffic management department. This will not only help the traffic management department to improve the management efficiency, but also effectively ease the urban traffic pressure, creating a smoother and safer travelling environment for the public.

References

[1] Kaili Z, Hailong Z, Jingyu L, et al. Traffic flow prediction based on graph convolutional neural network[J].Intelligent Computer and Applications, 2019.

[2] Fix. The design of the intelligent traffic management information system [J]. Integrated circuit applications, 2024, 9 (02): 298-299.

[3] Hu Xiaoyong, Zhang Jianjun, Yang Yun Hui, et al. Based on the nonlinear model predictive control of urban road network traffic flow optimization [J]. Journal of modern electronic technology, 2023, 46-48 (20)

[4] Li C, Ren X, Zhang Q, et al.Short-term Traffic Flow Prediction Based on Improved Deep Echo State Network[J].IEEE[2024-04-25].

[5] YAO Xu. Research on Forecasting method of urban Road Traffic demand in winter in Cold region [D]. Harbin Institute of Technology, 2014.

[6] Janessa Yin. Based on moderate traffic impact assessment of traffic simulation study [D]. Shenzhen University, 2021.

[7] Xiao Pei cheng, Cao Yang, Shen Qinqin, et al. Short-time traffic flow prediction by spatio-temporal graphical convolutional network based on ARMA filter [J/OL]. Computer Engineering and Applications:1-7[2024-04 05]

[8] XIA Jin, WANG Zhengqun, ZHU Shiming. A traffic flow prediction model based on time series decomposition [J]. Computer Applications, 2023, 43(04):1129-1135.

[9] Liu Fan. Several studies about skewed distribution [D]. Chang'an University, 2017. [10] Li Y, Zhang Chunxia. Hyperparameter estimation of random forest algorithm based on out-of-bag samples [J]. Journal of Systems Engineering, 2011, 26(04):566 572.