

A Study on Predicting Traffic Flow by Random Forest Based on Treatment of Eliminating Skewed Distribution

Yujia Zhao^{1, #}, Yang Yang^{2, #}, Xin Wen^{3, #, *}

¹School of Mathematical Sciences, Guizhou Normal University, Guiyang, China, 550025

²School of Mathematical Science and Artificial Intelligence, Dongbei University of Finance and Economics, Dalian, China, 116025

³Faculty of Economics and Management, Shanghai University of Applied Technology, Shanghai, China, 200030

* Corresponding Author Email: 15268995172@163.com

#These authors contributed equally.

Abstract. With the accelerating urbanization and complexity of transportation networks, short-term traffic flow prediction has become increasingly crucial in traffic management. Addressing this issue, this study introduces a random forest model for traffic flow prediction, incorporating deskewing treatment. Quantile-Quantile Plot analysis reveals the right-skewed distribution of traffic flow data, necessitating the use of a logarithmic function for deskewing. Further analysis indicates temporal patterns in the data. To validate the deskewing effect and model performance, Lasso, SVR, and Random Forest regression models were employed. The results demonstrate that deskewing significantly enhances prediction accuracy, with the Random Forest model outperforming others at various junctions.

Keywords: Traffic Flow, Skewed Distribution, Lasso, SVR, Random Forest Model.

1. Introduction

With the acceleration of the urbanization process and the increasing complication of traffic networks, short-term traffic flow prediction plays an increasingly important role in traffic management and control. The industry defines the prediction of traffic flow in the future five minutes to an hour as short-term traffic flow prediction. Accurate a short-term traffic flow prediction helps traffic management departments to formulate effective traffic evacuation strategies in advance, thus reducing traffic congestion and traffic accidents and improving the efficiency of urban traffic operations.

Fu Gui [1] proposed a short-time traffic flow prediction model based on support vector machine regression, which solved the linear regression problem by introducing a kernel function, and the experimental results showed that its prediction results were accurate, and the error was lower than that of the traditional Kalman filter method. Cheng Zheng [2], on the other hand, used the random forest model, which has strong generalization ability, convenient parameter adjustment, efficient computation and good stability, and the prediction accuracy is as high as 94%, which outperforms the support vector machine. Xing Panpan [3] combines support vector machine and neural network to construct a combined model to improve the accuracy of traffic flow prediction.

In summary, short-term traffic flow prediction is affected by complex and variable factors, and traditional prediction methods are difficult to achieve ideal results. With the continuous improvement of machine learning and data mining technology, the combination of the two has become an important research direction for traffic flow prediction. Therefore, this paper uses machine learning methods to predict short-term traffic flow, and adopts logarithmic function to deskew the right-biased data, so as to optimize the data normality and provide a more accurate data base for the subsequent prediction model. By comparing the three regression models of Lasso, SVR and RandomForestRegressor, this paper evaluates the performance of each model in the prediction of traffic flow data. The results show that the Random Forest model with unbiased processing has the best performance in traffic flow

prediction, which provides a new idea and method for traffic flow prediction, and provides a scientific basis for traffic management and scheduling.

2. Traffic Flow Data Exploration

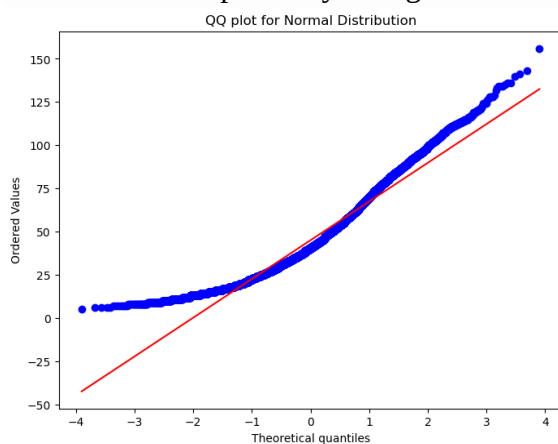
The data for this study was obtained from the address "<https://www.kaggle.com/>". It includes information such as Date Time, Junction, Vehicles, and ID. The dataset comprises 48,120 samples, capturing hourly traffic flow at four different intersections. Partial data is presented in Table 1 below:

Table 1. Partial presentation of traffic flow data

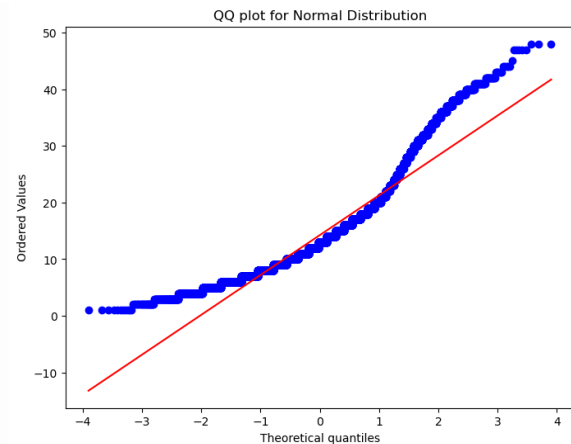
Date Time	Junction	Vehicles	ID
2015/11/1 0:00	1	15	20151101001
2015/11/1 1:00	1	13	20151101011
2015/11/1 2:00	1	10	20151101021
2015/11/1 3:00	1	7	20151101031
...			
2017/6/30 23:00	4	12	20170630234

2.1. Distribution of Traffic Flow

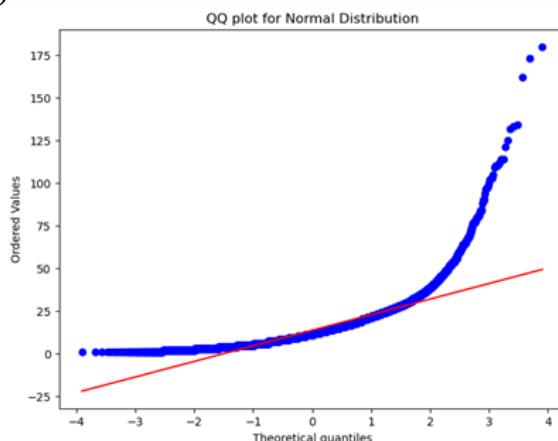
The traffic flow data of Junction1-Junction4 in Table 1 above are plotted using Quantile-Quantile Plot, demonstrated respectively in Figure 1 below:



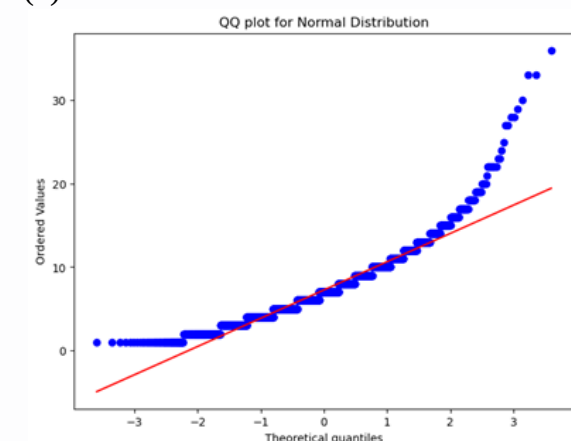
(a) Distribution of traffic flow data in Junction 1



(b) Distribution of traffic flow data in Junction 2



(c) Distribution of traffic flow data in Junction 3



(d) Distribution of traffic flow data in Junction 4

Figure 1. QQ Plot of traffic flow data in Junction1-Junction4

The skewed distribution is a concept opposite to normal distribution, whose image shows the characteristics of left-right asymmetry. It can be used to reflect the distribution state of a continuous variable. The red line in Quantile-Quantile Plot indicates the normal distribution and the blue scatters indicate the sample data. The closer the blue scatters are to the red reference line, the more it conforms to the normal distribution. Observing (a)-(d) in Figure 1 above, it is found that the data present the following characteristics: the traffic flow data at the two ends deviate from the red reference line to a greater extent and distribute above the reference line, while the data in the middle part deviate to a lesser extent and distribute more centrally. Therefore, it is inferred that the data show the characteristics of skewed distribution, and the data distribution presents in the figure belongs to the right-skewed distribution.

Machine learning algorithms usually assume that the data obey a normal distribution, because based on a normal distribution and under feature engineering, its data can extract more effective information and is more beneficial to subsequent research [4]. Therefore, the traffic flow data above needs to be deskewed.

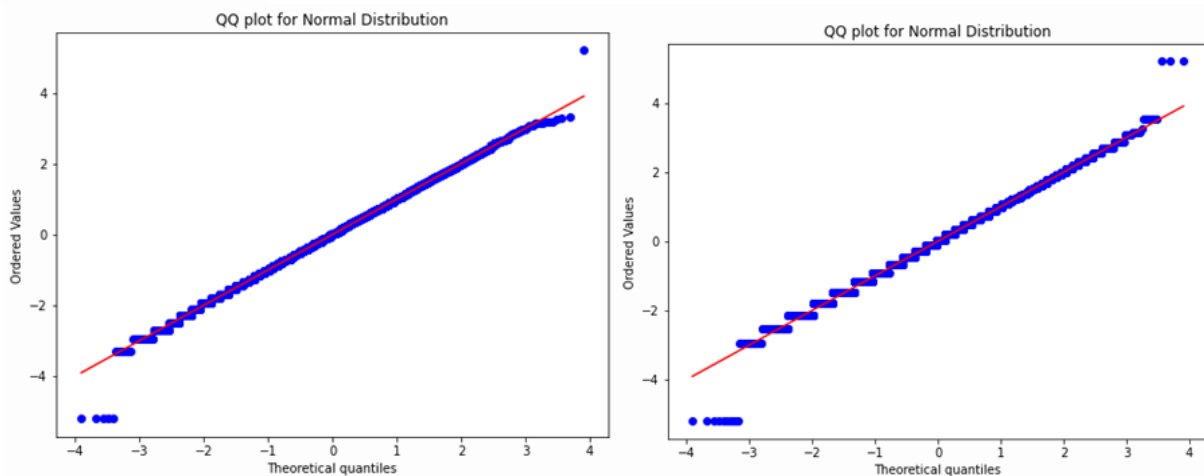
2.2. Deskewing Treatment Method

For a logarithmic function with a base of more than one, its function value increases as the independent variable increases, but the increase range decreases gradually, which means its derivative monotonically declines. So the function convert has the effect of shifting large data to the left, leading to bringing the data distribution closer to a normal distribution. Bring the original traffic flow data x into $f(x) = \ln x$, and a set of datasets of value domain is obtained. Use this set of data to replace the original traffic flow x , thus making the original skewed distribution of traffic flow converge to the natural distribution and satisfying the need for machine learning algorithms with requiring data to be normally distributed. The formula used is as follows:

$$f(x) = \ln x \quad (1)$$

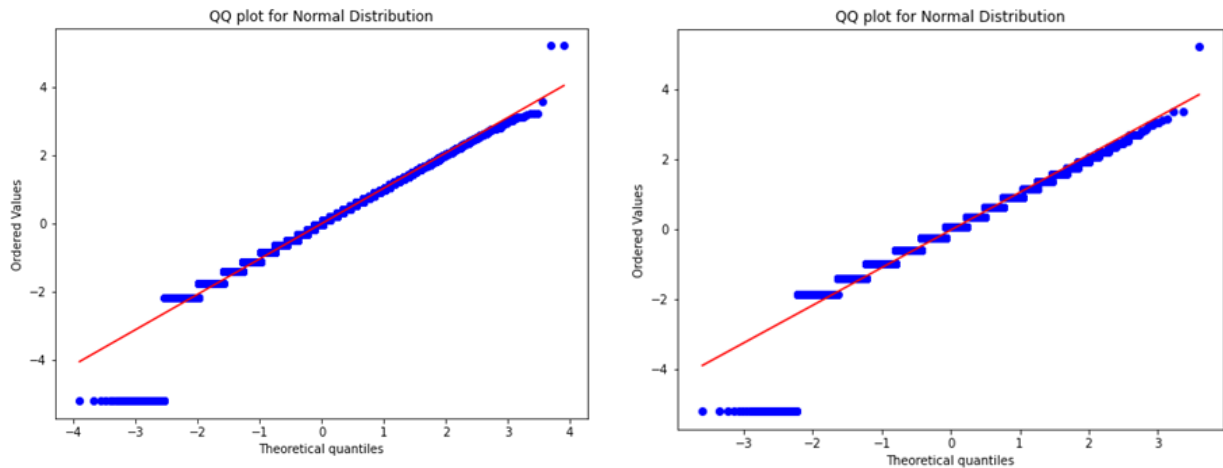
Among the formula, $f(x)$ represents the log-transformed traffic flow value, and x represents the original traffic flow value.

The traffic flow data distribution obtained after deskewing processing is shown in Figure 2. As can be seen from the figure, the blue scatters in the transformed Quantile-Quantile Plot are close to the red reference line, which proves that the traffic flow data after deskewing processing is close to the normal distribution.



(a) Distribution of traffic flow data at Junction 1 after deskewed treatment

(b) Distribution of traffic flow data in Junction 2 after deskewed treatment



(c) Traffic flow data distribution of Junction 3 after deskewed treatment

(d) Distribution of traffic flow data in Junction 4 after deskewed treatment

Figure 2. QQ Plot of traffic flow data after deskewed in Junction1-Junction4

At the same time, the predicted traffic flow data are inversely converted by the exponential function in this study, which ensures the originality of the data. The formula used is as follows:

$$x = 10^{f(x)} \quad (2)$$

2.3. Time Law Analysis of Traffic Flow

Arranged according to time series, the distribution of traffic flow exists a certain law of change over time. To facilitate the exploration of the law, the time series distribution of traffic flow at four different intersections is plotted below.

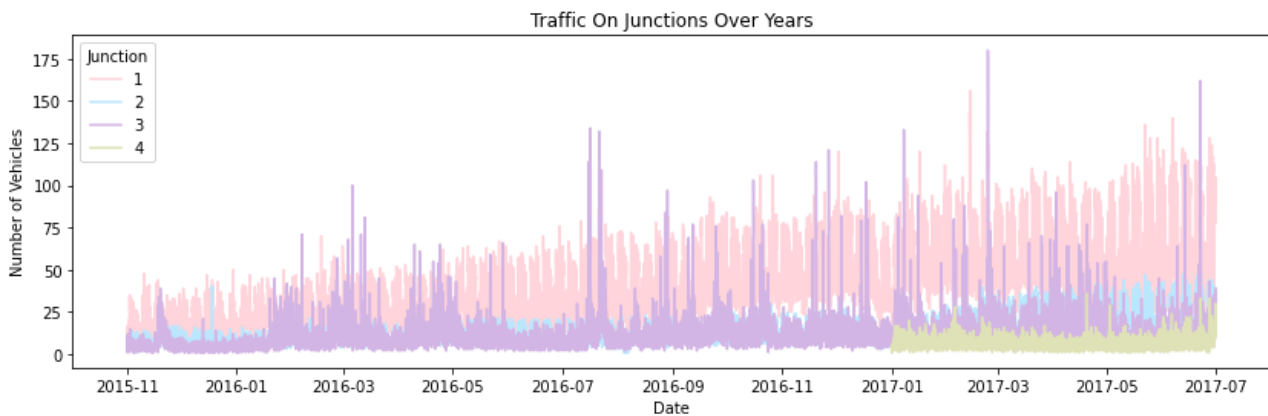


Figure 3. Time series distribution of traffic flow

Observing Figure 3 above, it is found that all four intersections show an increasing trend in traffic flow year on year. Among them, it reveals that the data related to intersection 4 is found missing between November 2015 and January 2017, but it also remained unchanged in terms of the overall yearly increasing trend.

To observe the traffic flow in different time states in a more detailed way, this study visualizes and analyzes the state of traffic flow over time at each intersection during the review period. The time series is also specifically expanded into year, month, day, and hour, and the detailed results are shown in Figure 4:

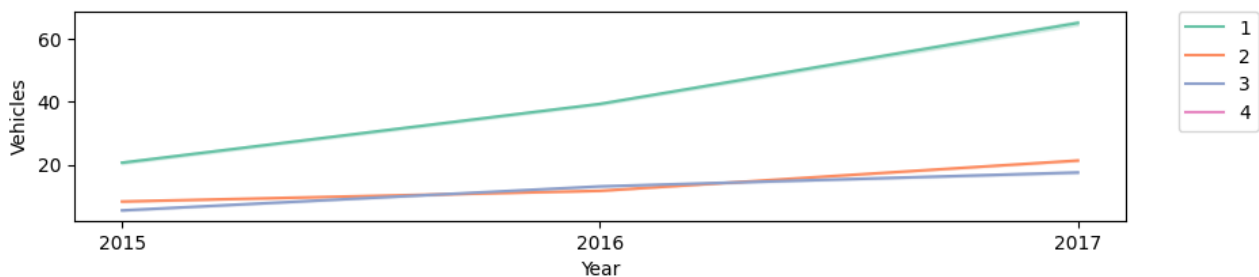
(1) As shown in Figure (a), the data for all intersections except Intersection 4 show an upward trend. Intersection 1 has the most pronounced upward trend, and Intersection 2 and Intersection 3 have similar upward trends. Consistent with the observations made above of Figure 3, the data of Intersection 4 is limited and does not span more than one year. Intersection 1 has the highest traffic

volume and is presumed to be in a busy area.

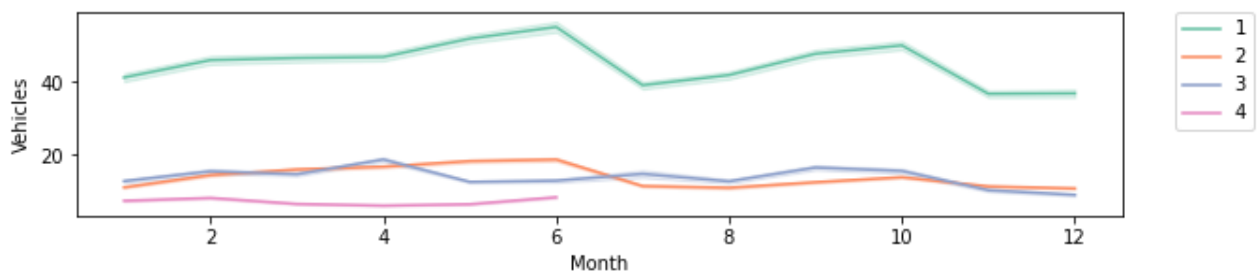
(2) As shown in Figure (b), before and after June, the traffic flow at all intersections except Intersection 3 have obvious ups and downs, which is presumed that they have some correlation with summer vacations and vacation-related activities. From a quarterly perspective, the traffic flow data presentation status of all intersections remains consistent.

(3) As shown in Figure (c), the average daily flow fluctuates the most at Intersection 1 and the least at Intersection 2. During one week, the traffic is smoother and lower on weekends, while the change in traffic flow is stabler from Monday to Friday.

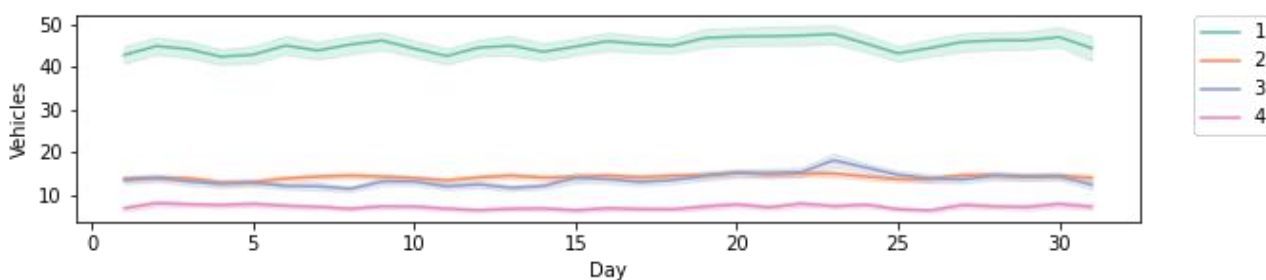
(4) As shown in Figure (d), during a whole day, there is a sustained peak from 10:00 a.m. to 8:00 p.m. and then a drop to a trough around 5:00 a.m., which is consistent with common sense. In this case, the traffic flow at Intersection 1 fluctuates greatly, with high traffic volumes and obvious peak and valley.



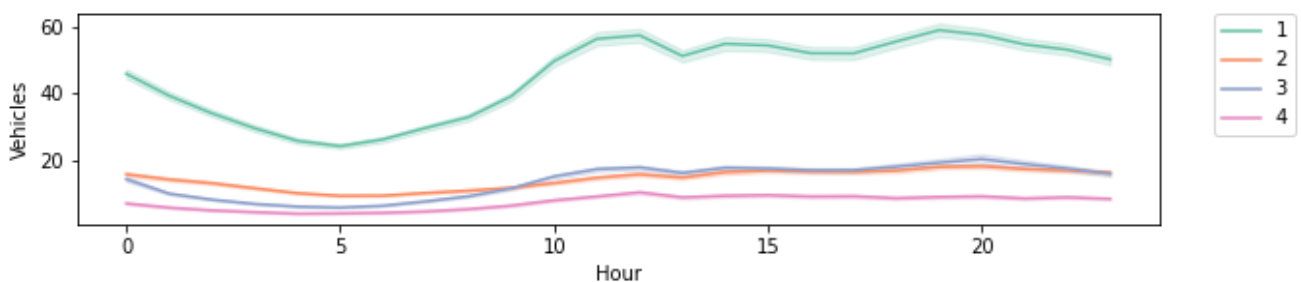
(a) Yearly trend of traffic flow data



(b) Monthly trend of traffic flow data



(c) Daily trend of traffic flow data



(d) Hourly trend of traffic flow data

Figure 4. Trend of traffic flow data

2.4. Heat map of correlation coefficients

Correlation Coefficients can be used to measure the degree of linear correlation between two variables [5-6]. The formula for calculating the correlation coefficient r between variable X and variable Y is as follows:

$$r_{x,y} = \frac{CoV(X,Y)}{\sqrt{Var[X]Var[Y]}} \quad (3)$$

$CoV(X, Y)$ is the covariance of X and Y , $Var[X]$ and $Var[Y]$ are the variances of X and Y , respectively. The correlation coefficient r is in the interval $[-1,1]$: when $r > 0$, the two variables are positively correlated; when $r < 0$, the two variables are negatively correlated; when $r = 0$, the two variables are not linearly correlated. It is generally accepted that when $|r| > 0.8$, the two variables are strongly correlated, while $0.5 < |r| < 0.8$ indicates moderately correlated, $0.3 \leq |r| < 0.5$ means weakly correlated and $|r| < 0.3$ embodies that they are almost uncorrelated.

To explore the main factors affecting traffic flow, this study calculates the Pearson correlation coefficient matrix between intersection, year, month, day, hour, etc., and traffic flow. To offer convenience for observation, the correlation matrix is visualized, and the correlation coefficient heat map is finally received as shown in Figure 5 below:



Figure 5. Heat map of correlation coefficient of traffic flow time series

Observing Figure 5 above, it is found that there is a strong correlation as high as -0.61 between Junction and traffic flow. Weak linear correlation exists between the rest of the features and traffic flow. Therefore, the subsequent model-establishing can be considered to use a nonlinear model.

3. Establishment of Model

3.1. Data Segmentation

Under normal circumstances, the data will be divided into train sets and test sets when exploring models related to machine learning. The train set is used to fit the data and train the model; the test set is used to compare the original data and the fitting results in order to evaluate the model performance and validate the model fitting effect. A total of 48,120 data samples are used in this study, and the original data set are divided to improve the accuracy and generalization ability of model-building. Due to the large sample set in this study, the original data are randomly divided into a train set and a test set according to 8:2.

3.2. Grid Search and Cross Validation

Hyperparameters are parameters that must be set before machine learning can be performed. The selection of these parameters is crucial for model building. Optimizing hyperparameters can significantly enhance model performance. The grid search method is an exhaustive search for specifying parameter values to find the best-performing parameters. In this study, this method is used with 5-fold cross-validation to search multiple training and test sets, calculate the average model performance, and finalize the hyperparameter selection. This process aims to optimize the parameter combinations.

3.3. Model Hyperparameter

In this study, Lasso, SVR, and random forest, three regression models are used to predict traffic flow data respectively [7-9]. According to the results of grid search and cross validation, the hyperparameters of each model are selected as follows in Table 2. The hyperparameters not listed are default values:

Table 2. Model hyperparameters

Mould	Parameter	Parameter value
Lasso	alpha	0
SVR	C	100
Random forest	min_samples_leaf_estimators	3

3.4. Model evaluation

The above table presents the predictive evaluation indicators of the cross validation set, the train set, and the test set, and measures the prediction effect of the random forest through quantifying indexes. Among them, the evaluation indicators of the cross validation set can continuously adjust the hyperparameters to obtain a reliable and stable model.

MSE (Mean Square Error): The squared expected value of the difference between the predicted value and actual value. The smaller the value, the more accurate the model.

MSE (Root Mean Square Error): The square root of the MSE. The smaller the value, the more accurate the model.

MAE (Mean Absolute Error): The average value of the absolute error, which reflects the actual situation of the prediction error. The smaller the value, the more accurate the model.

R²: Comparing the predicted value to the case where only the average value is used, the closer the result is to 1, the more accurate the model.

Table 3. Value of model assessment metric for Junction 1 without deskewed treatment

Mould	MSE	RMSE	MAE	R ²
Lasso	155.51	12.47	10.00	0.70
SVR	42.23	6.50	4.57	0.92
Random forest	19.40	4.40	3.10	0.96

Table 4. Value of model evaluation index for Junction 1 after deskewing treatment

Mould	MSE	RMSE	MAE	R ²
Lasso	143.59	11.98	9.26	0.73
SVR	32.45	5.70	3.98	0.94
Random forest	19.59	4.42	3.08	0.96

According to the above table 3 and table 4 above, it is found that the lasso linear regression model predicts poor results, it can be initially inferred that the relationship between features and traffic flow is nonlinear. So it is necessary to use a nonlinear model for prediction. Then Junction 2, 3, and 4 are used a nonlinear random forest model, and the prediction results are as follows in Table 5:

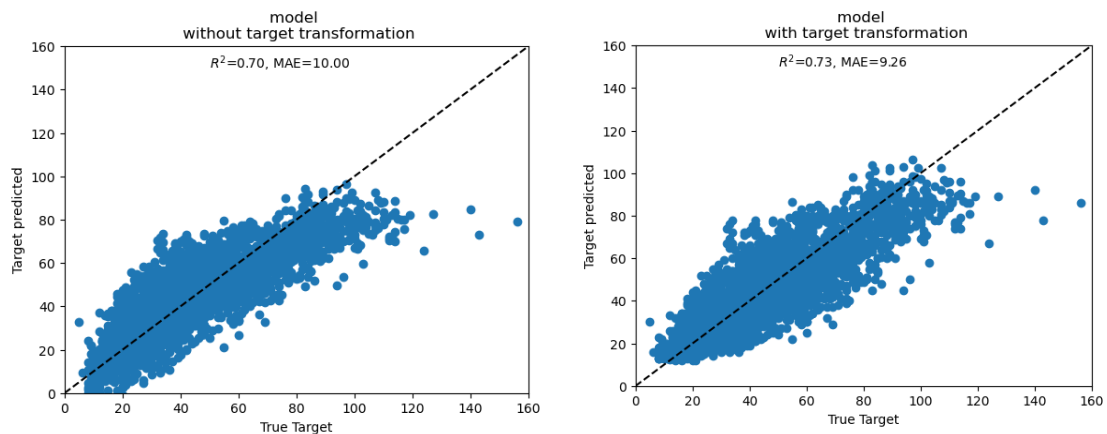
Table 5. Model evaluation metric for Junction2, 3, and 4 without deskewed treatment

Intersection (of roads)	MSE	RMSE	MAE	R ²
Junction 2	5.21760	2.2842	1.75	0.91
Junction 3	34.1635	5.8450	2.98	0.71
Junction 4	5.6445	2.3758	1.82	0.51

Table 6. Model evaluation metric for Junction2, 3, and 4 after deskewing treatment

Intersection (of roads)	MSE	RMSE	MAE	R ²
Junction 2	5.3486	2.3127	1.7400	0.9100
Junction 3	34.5095	5.8745	2.8300	0.7000
Junction 4	5.9757	2.4445	1.8300	0.4700

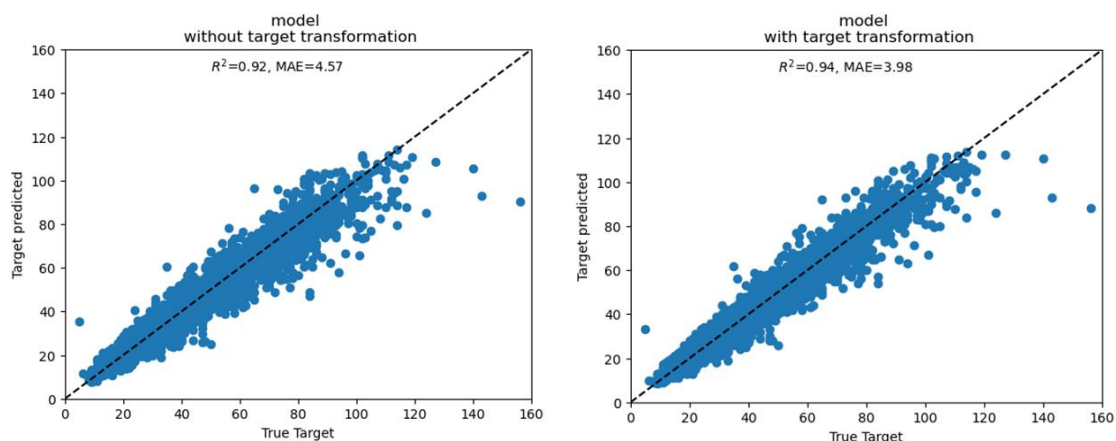
3.5. Model results



(a) Without deskewed treatment

(b) After deskewing treatment

Figure 6. Traffic flow prediction result of Lasso for Junction 1



(a) Without deskewed treatment

(b) After deskewing treatment

Figure 7. Traffic flow prediction result of SVR for Junction 1

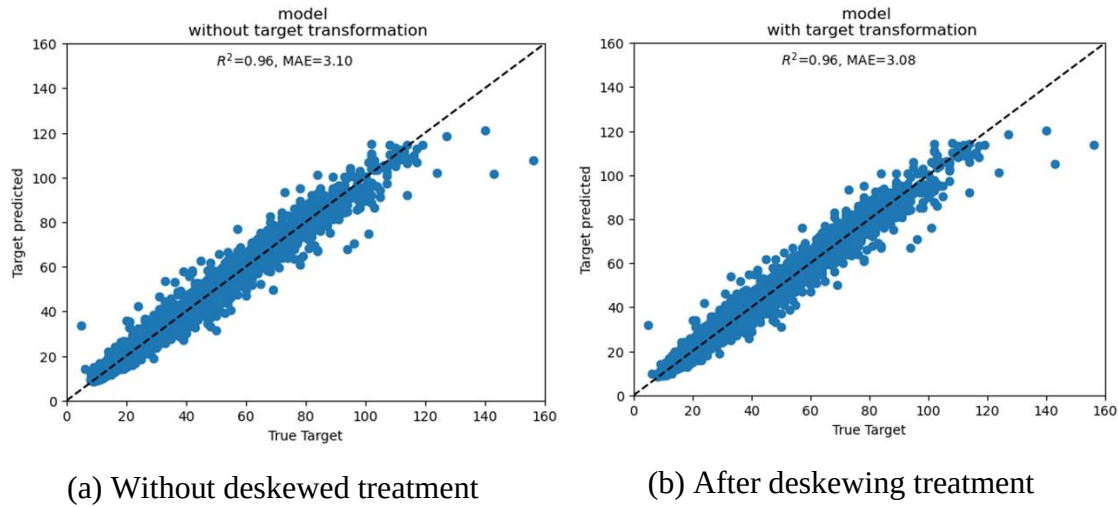


Figure 8. Traffic flow prediction result of random forest for Junction 1

Analysis of Figures 6-8 reveals that the Lasso model, without deskewing, achieved $R^2=0.70$, $MAE=10.0$, and $MSE=155.5094$, while the SVR model achieved $R^2=0.92$, $MAE=4.57$, and $MSE=42.2279$. The random forest model outperformed both, with $R^2=0.96$, $MAE=3.10$, and $MSE=19.4003$. Upon deskewing, the Lasso model improved to $R^2=0.73$, $MAE=9.26$, $MSE=143.5884$, while the SVR model reached $R^2=0.94$, $MAE=3.98$, $MSE=32.4510$. The random forest model, at $R^2=0.96$, $MAE=3.08$, $MSE=19.5905$, remained superior. This verifies the conjecture of deskewing enhancing model performance, particularly for nonlinear models [10-11].

Figures (a)-(c) in Figure 9 below show the predicted traffic flow results for Junction 2-Junction 4:

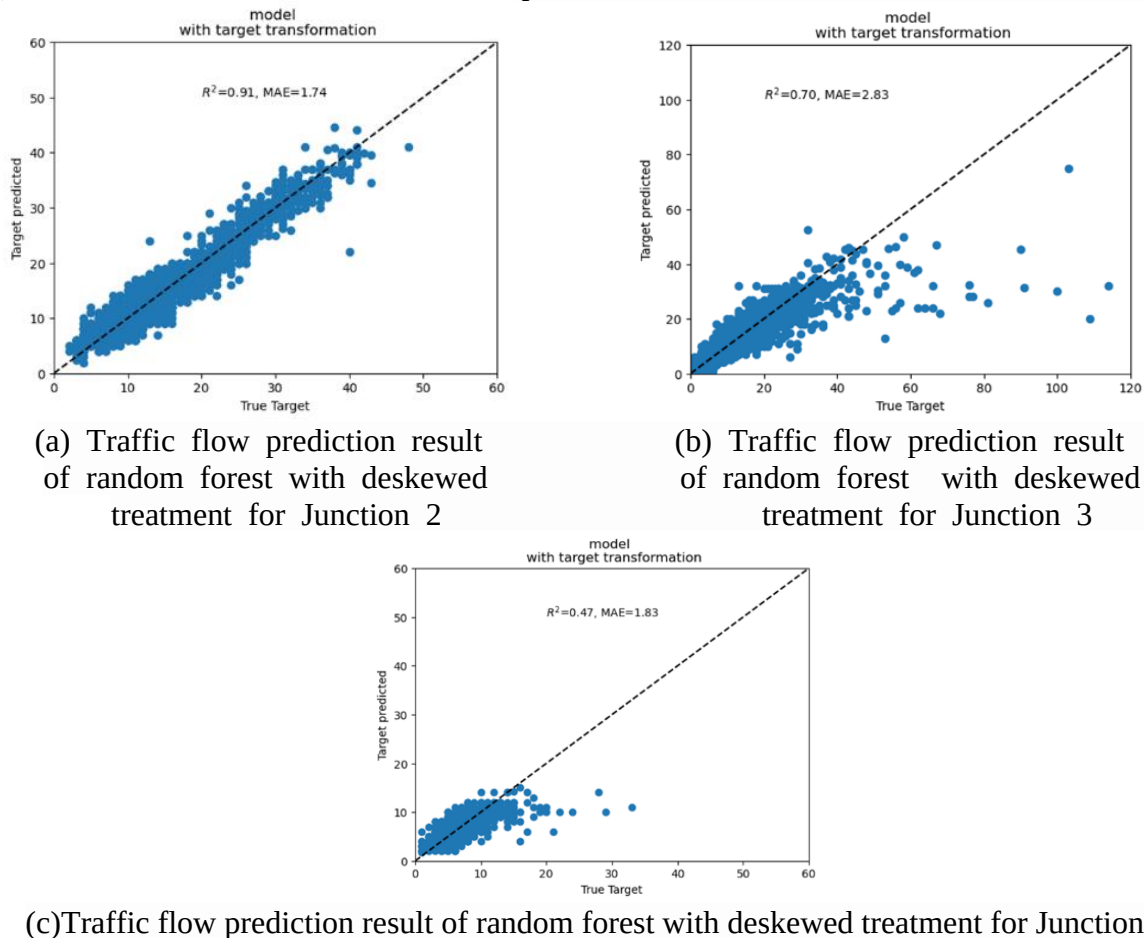


Figure 9. Random forest prediction result for Junction 2-Junction 4

The prediction result indicates that: the random forest model with deskewed treatment for Junction

2 has an R^2 value of 0.91, MAE value of 1.74 and MSE value of 5.3486; the random forest model with deskewed treatment for Junction 3 has an R^2 value of 0.70, MAE value of 2.83 and MSE value of 34.5095; the random forest model with deskewed treatment for Junction 4 has an R^2 value of 0.47, MAE value of 1.83 and MSE value of 5.9757. The predictive result of the three intersections are relatively impressive, so the models are correctly chosen, the random forest regression model is the best.

4. Conclusion

In this paper, we propose a random forest prediction model for traffic flow, incorporating skewed data processing, and conduct prediction simulations using existing historical traffic flow data. Firstly, Quantile-Quantile (Q-Q) Plot analysis reveals that the traffic flow data exhibit a right-skewed distribution, prompting the application of a logarithmic transformation to normalize the data. Secondly, the traffic flow data display temporal patterns varying with year, month, day, and hour. Pearson correlation analysis identifies a weak linear correlation between feature values and traffic flow. Finally, we utilize three regression models—Lasso, SVR, and random forest—to predict the traffic flow data. The results indicate that predictions based on deskewed data are more accurate than those without treatment, with the nonlinear random forest model proving most effective.

Real traffic flow forecasts can be influenced by various factors. For instance, weather and social events, such as holidays, can cause significant fluctuations in traffic flow. Therefore, the next step is to introduce additional impact factors to further enhance the model's prediction accuracy, thereby providing crucial support for traffic management.

References

- [1] FU Gui, KAO Qiang, LU Feng, XU Zixin. A short-time traffic flow prediction model based on support vector machine regression[J]. Journal of South China University of Technology(Natural Science Edition), 2013, 41(09): 71-76.
- [2] CHEN Zheng, CHEN Xianfu. A short-time traffic flow prediction method based on random forest model [J]. Microcomputer and Applications, 2016, 35(10): 4.
- [3] Xing Panpan, Li Chunmei.svm and neural network combination model prediction analysis [J]. Electronic technology, 2023, 52 (12): 387-389.
- [4] ZHAO Xinbin, WANG Yan. A flight quality risk measure based on skewed distribution [J]. China Journal of Safety Science, 2019, 29(10): 160-166.
- [5] JI Deyang, JIN Feng, DONG Lei, et al. Data restoration of photovoltaic power plants based on Pearson's correlation coefficient[J]. Chinese Journal of Electrical Engineering, 2022, 42(04): 1514-1523.
- [6] Yuan-Shang Zhao, Wei-Fang Lin. Study of typical scenarios based on Pearson correlation coefficient fusion of peak density and entropy weight method [J]. China Electric Power, 2023, 56(5): 193-202.
- [7] TIAN Donghai, CAI Qiquan, HU Junping, et al. Analysis of main control factors of gathering pipeline based on LASSO association rule algorithm [J]. Oil and Gas Field Surface Engineering, 2024, 43(04): 41-47.
- [8] CHENG Shan-Ying. Construction of short-time traffic flow prediction model based on SVR [J]. Information and Computer, 2022, 34(14): 18-20.
- [9] WANG Huiru, FENG Xuejun, SHEN Jinxing, et al. Carbon emission peak prediction of transportation industry in Jiangsu Province based on SVR [J]. Environmental Science and Technology, 2022, 35(03): 58-63.
- [10] XU Xiujuan, BAI Yulin, XU Lu, et al. Traffic flow prediction based on random forest algorithm under severe weather condition s[J]. Journal of Shaanxi Normal University(Natural Science Edition), 2020, 48(02): 25-31.
- [11] Li Xiuzhen. Research and application of intelligent highway traffic accident prediction based on random forest model [J]. Shanxi Transportation Science and Technology, 2023(5): 103-105.