Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/140/2025.21273

Analysis Model Integrating Traffic Volume and Temperature for Traffic Flow Prediction

Yishan Liu[1,a,*]

[1] *College of Metropolitan Transportation, Beijing University of Technology, Beijing, China*
a. liuyishan@emails.bjut.edu.cn
**corresponding author*

Abstract: Rising global vehicle numbers lead to traffic congestion, which is influenced by temperature. Nowadays, temperature forecasting has reached a high level of sophistication, so it is feasible to predict traffic volume based on the forecasting temperatures. To predict the traffic volume, collected the traffic and weather data of I-94 Interstate highway. Based on the data, the simple linear regression model (LR), polynomial regression model (PR) and random forest (RF) were applied. With the comparison between simulation performances, RF demonstrated the highest accuracy, with a correlation coefficient (R2) of 0.8364. To bolster the simulation performances of regression models, more constraints should be involved such as rainfall, snowfall, and cloud cover, should also be considered. Besides, more sophisticated algorithms, such as regression-enhanced random forests (RERFs) and improved random forest classifiers (RFC), should be applied as well. In conclusion, as more related factors are involved, the performance of regression models will be better, which serve for traffic management effectively, with accurate traffic flow prediction.

Keywords: traffic volume, temperature, regression model, simulation.

1. Introduction

As the global economy progresses, the number of motor vehicles increases sharply. In 2022, there were about 1.446 billion cars in the world [1]. Obviously, people are dependent more and more severely on motor vehicle travel. In the United States, 91% of adults commute to work using personal vehicles [2]. Consequently, the road traffic volume exploded in growth, leading to numerous problems in traffic systems, such as traffic congestion. An effective traffic flow prediction serves as a vital foundation for transportation management, forestalling traffic problems. The key to achieving an accurate and reliable traffic flow prediction lies in modeling the complex and dynamic correlations among impact factors [3]. Thus, to predict traffic volume precisely, the regression models between impact factors and traffic volume have been established. A multitude of scholars conducted comprehensive studies on regression models for traffic flow forecasting. Bayati et al. developed a Gaussian process regression ensemble model for network traffic prediction, with 12% average prediction accuracy for different datasets [4]. In order to overcome the problems of low accuracy and the time-consuming of traditional prediction methods for short-term traffic flow in urban, Li proposed prediction methods based on a multiple linear regression model [5]. Yu et al. employed prediction models combining the Gorilla Optimization Algorithm and Whale Optimization Algorithm respectively, and the results demonstrated satisfactory training and prediction capabilities [6]. Huang et al. developed a hybrid model of a neural network with VMD-CNN-GRU for traffic flow prediction, which significantly raises the precision of traffic flow forecasting [7]. Yu identified different patterns of traffic flows by a negative binomial model with smoothing splines [8]. Therefore, predicting traffic flow based on regression models is practical, and there are apparent research gaps concerning the specific impacts of temperature on traffic flow and the relationship between the two. However, climate change has been a global threat, particularly manifesting in global warming [9, 10]. As temperature changes, the life behaviors of people are transforming consequently [11]. Temperature is one of the key factors influencing transport patterns to people, affecting their decisions through multiple dimensions such as comfort, safety, health, and economy [12]. Nowadays, temperature forecasting has reached a high level of sophistication, so it is feasible to predict traffic volume based on the forecasting temperatures [13]. Therefore, to accurately predict traffic volume, temperature should be included as a significant factor affecting traffic volume.

In this study, to predict traffic flow, a linear regression model (LR), polynomial regression model (PR) and random forest (FR) model were built. Besides, to find the optimum model, according to mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE), and correlation coefficient (R2) as simulation performances, the prediction accuracy of models were compared.

2. Data and Method

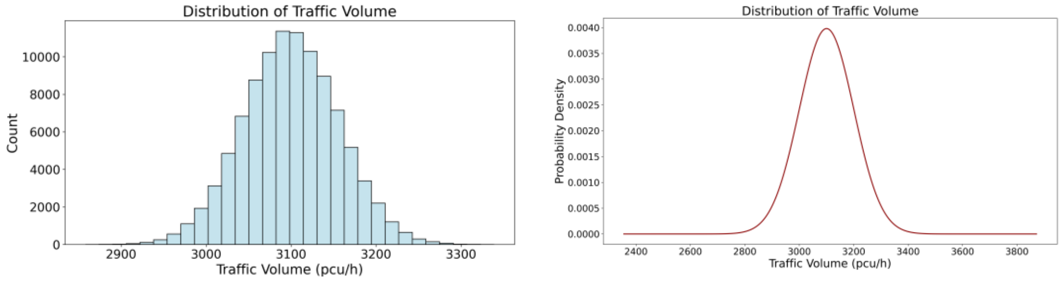
2.1. Data Collection and Description

To obtain a wider numerical range about temperature, the I-94 Interstate Expressway in Indiana, where the seasons are distinct, was selected as the subject of traffic volume investigation. Information on daily hourly traffic volume, temperature, rainfall, snowfall, cloud cover, and visibility on the I-94 Interstate highway was collected from 2012 to 2013. To minimize the error of the regression model possibly, selected 3 months with significant temperature differences: October 2012, January 2013, and May 2013. The details of the differences in these months are shown in Table 1.

Table 1: Temperature of three Months comparing.

	Maximum Temperature (K)	**Minimum Temperature (K)**
October	290	265
January	276	240
May	296	281

The occurrences of different traffic volumes every hour in survey days were counted, and the distribution of traffic volume was shown in Figure 1(a), and the normal distribution curve was shown in Figure 1(b).



(a) (b)

Figure 1: Investigation Results. (a) is the Distribution of Traffic Volume,(b) is the Normal Distribution Curve. (Picture credit: Original).

From Figure 1 (a) and Figure 1 (b), the traffic volume that appeared most frequently was approximately 3100 pcu/h, with a frequency of over 10,000 occurrences and a probability density of nearly 0.0040. Thus, the traffic conditions of the I-94 Interstate highway exhibited a high degree of smoothness and fluidity, indicating an obvious absence of traffic congestion. The basic information on daily hourly weather conditions during the survey period is shown in Figure 2.

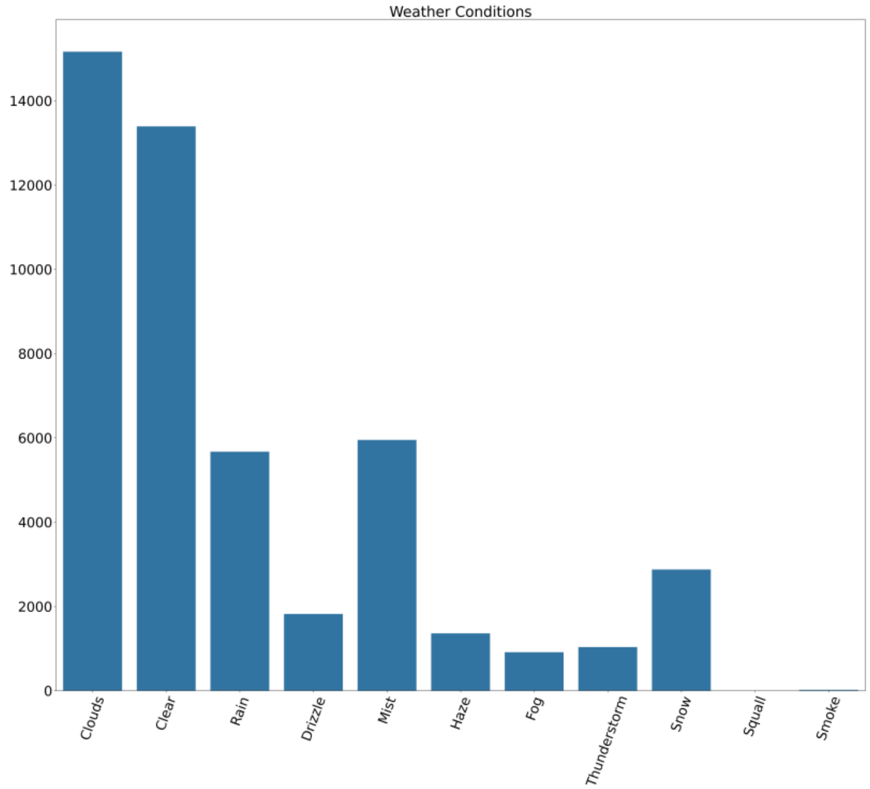


Figure 2: Weather Conditions. (Picture credit: Original).

From Figure 2, the weather that appeared most frequently secondly was clear, with a frequency of more than 14000 times, which illustrates that the visibility for drivers on the I-94 Interstate expressway was impressive. Therefore, based on the traffic conditions and weather conditions, the traffic data from the I-94 Interstate highway was reliable.

2.2. Data Pre-processing

A continuous 20-day period within a month was chosen as the subject, from October 2nd to 21st in 2012, from January 1st to 20th and from May 7th to 27th in 2013. The average traffic volume per hour of a day, and the average temperature per hour of a day, were calculated based on the original data. The processed data is shown in Table 2.

Table 2: Traffic Volume and Temperature in Three Months.

October		January		May	
Traffic Volume (pcu/h)	Temperature (K)	Traffic Volume (pcu/h)	Temperature (K)	Traffic Volume (pcu/h)	Temperature (K)
4219	290	1913	240	3838	288
4017	286	2556	258	3754	291

Table 2: (continued).

4098	289	3540	264	3777	291
4043	282	3402	263	3838	285
3257	278	3208	267	3987	284
2926	265	3185	264	3844	282
3422	277	3361	264	3605	281
3770	283	3364	271	3812	286
3904	281	3749	270	4847	296
3911	279	3372	271	4738	293
4026	281	3730	274	4741	294
3535	277	3762	276	3908	288
3852	283	3338	264	3929	289
3993	283	2668	258	4209	293
3528	271	2851	259	4550	291
3657	288	2301	244	3572	286
3683	285	3814	269	3903	283
3414	280	3165	262	3604	285
3732	282	3853	271	3473	286
3533	273	3202	267	3984	287

2.3. Linear Regression

Linear regression (LR) is a common machine learning algorithm used to analyze the linear relationship between independent variables (input features) and dependent variables (output results). The parameter of the random state was set as 2529. Based on the train set, established a simple LR in computer. By LR, the mathematical expression between traffic volume (V) and temperature (T) was shown in equation (1).

$$V = 35.77941027 \cdot T - 6278.653997651144 \tag{1}$$

2.4. Polynomial Regression

Polynomial regression (PR) is an extension of LR that models the non-linear relationship between the dependent variable (output features) and one or more independent variables (input features). PR captures non-linear relationships by transforming the input features into polynomial terms, such as squares and cubes of the original features, allowing for more complex patterns in the data. PR was established by the function of Linear Regression in the computer. By PR, the mathematical expression between V and T was shown in equation (2).

$$V = -2015.47 + 0.54 \cdot T - 0.07 \cdot T^2 \tag{2}$$

2.5. Random Forest

Random Forest (RF) is an ensemble machine learning algorithm that operates by constructing multiple decision trees during training time and outputting the average prediction in regression tasks, improving the prediction accuracy and control over-fitting. RF is used to analyze the complex relationships between independent variables (input features) and dependent variables (output results) by leveraging the wisdom of the crowd principle, where the combined predictions of multiple decision trees are more accurate and robust than those of a single tree. To establish RF in the computer, the number of decision trees used was set to 100, and the seed for the random number generator was set to 42. RF was a kind of algorithm, without a certain mathematical expression.

3. Results and Discussion

3.1. Experimental Setup

LR, PR, RF were established by the function in scikit-learn, without certain optimizers, learning rates, or loss functions respectively. To ensure the rigor of the experiment, training sets of LR, PR, RF account for 70% of the dataset, and the test sets account for 30%.

3.2. Experimental Results

The experimental results of 3 regression models were demonstrated by graphs, for LR, the prediction results were shown in Figure 3.

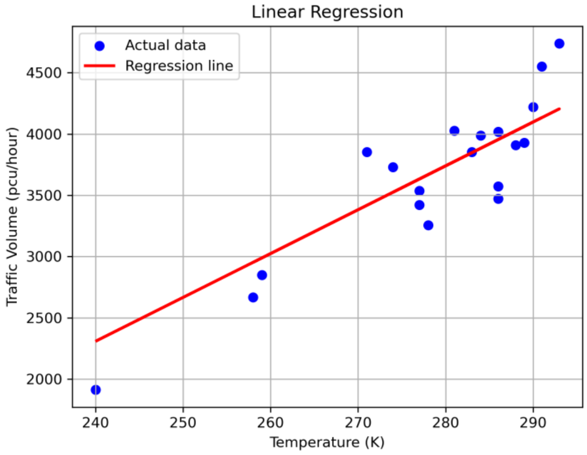


Figure 3: Linear Regression. (Picture credit: Original).

In Figure 3, the first predicted values in the initial stage were significantly higher than the actual value. Besides, the growing speed of V in the expression did not catch up with the speed of actual V, when the T was over 290K. Thus, there were obvious prediction errors in it when the T was remarkably high.

For PR, the prediction results are shown in Figure 4.

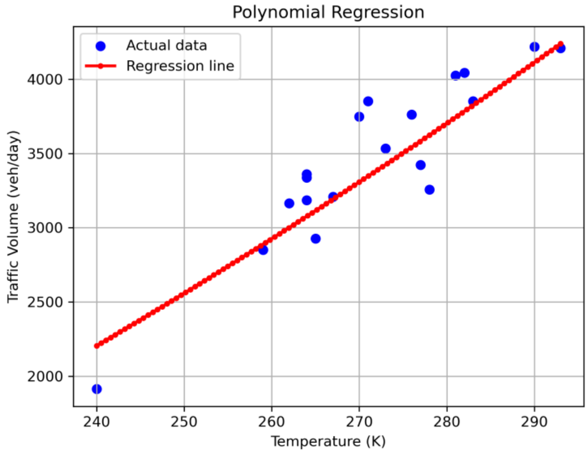


Figure 4: Polynomial Regression. (Picture credit: Original).

From Figure 4, the distribution of test sets is concentrated from 260 K to 290 K. The experimental values in the middle stages were pretty near to the actual value. Nevertheless, compared to LR, it showed that the growth rate of V in the equation was analogous to the actual growth rate of V in the later stages, and a reliable prediction of the maximum value of V. However, the prediction of the minimum value in the beginning is still imprecise in PR. Therefore, the accuracy of this mathematical expression increased in the final stages, which warranted a refinement.

For RF, the prediction results are shown in Figure 5.

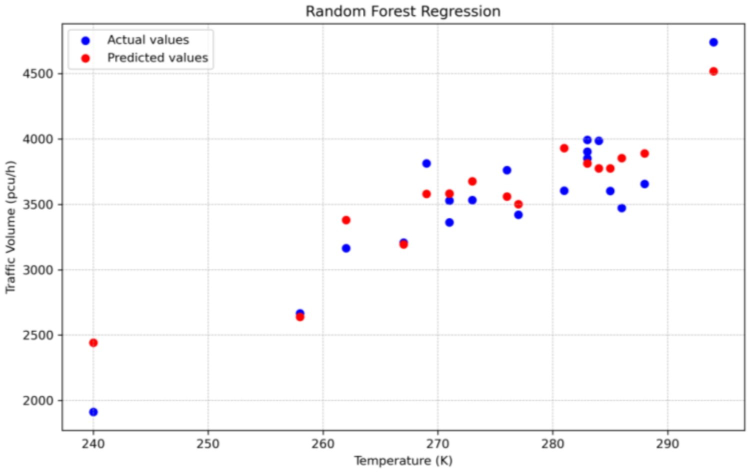


Figure 5: Random Forest Regression. (Picture credit: Original).

From Figure 5, in the initial stage, the two predicted values were extremely close to actual values. In addition, it was obvious that when the T was 258K, 267K, 271K, 273K, 277K, and 283K, the prediction value of V was near to the actual value of them, even nearly equaled to the actual value at 258K. However, the prediction was imprecise, higher than the actual value when T was low and lower than the actual value when T was high.

To analyze the prediction accuracy of the 3 regression models more deeply, compared their MAE, MAPE, RMSE, and R2 as simulation performance metrics. The details are shown in table 3.

Table 3: Simulation Performance Metrics Comparing.

	MAE	**MAPE**	**RMSE**	**R2**
LR	252.8177	7.33%	296.4829	0.7863
PR	203.1246	6.08%	245.0664	0.7972
RF	187.9163	5.83%	225.9994	0.8364

LR possessed the highest MAE, MAPE, RMSE, and lowest R2 which indicated that LR was inefficient in prediction. PR demonstrated the secondary lowest MAE, MAPE, and RMAE, and the secondary highest R2, which inferred that PR was precise comparatively in prediction. RF showed the lowest MAE, MAPE, and RMSE, with the highest R2, which illustrated that RF was optimum to describe the relationship between V and T, and predict V.

LR, PR, RF all exhibited varying prediction accuracy in certain degrees, each with a MAPE less than 10%, respectively. However, none of them predict an extremely low value of traffic volume well. Besides, none of their R2 values exceeded 0.9, with the maximum value being approximately 0.8364, which was still below 0.85. Thus, the simulation performances of these models were inadequate, especially for extreme values. To improve the performances, it was necessary to involve more weather impact factors and constraints in regression models, such as rainfall, snowfall, and cloud cover. Furthermore, it was considerable to apply more sophisticated algorithms in prediction, such as regression-enhanced random forests (RERFs) and improved random forest classifier (RFC).

4. Conclusion

To predict V accurately, the regression models between V and T. LR, PR, and RF were developed, and each of them demonstrated a mathematical relationship between V and T. LR, PR, and RF all exhibited varying prediction accuracy in certain degrees. According to the prediction results, the simulation performances of these models were inadequate for extreme values. Based on simulation performance metrics, involving MAE, MAPE, RMSE, and R2, FR showed the optimum simulation performance. The simulation performances of regression models were rough, with insufficient weather impact factors and constraints considered in the models. For future research outlook, it was considerable to consider impact factors, such as rainfall, snowfall, and cloud cover, and apply more sophisticated algorithms in prediction, such as RERFs and RFC.

References

[1] *Zahoor A, Mehr F, Mao G, et al. (2023). The carbon neutrality feasibility of worldwide and in China's transportation sector by E-car and renewable energy sources before 2060[J]. Journal of Energy Storage, 61: 106696.*

[2] *Moody J, Farr E, Papagelis M, et al. (2021). The value of car ownership and use in the United States[J]. Nature Sustainability, 4(9): 769-774.*

[3] *Li X, Yin X, Huang X, et al. (2024). Multi-dynamic residual graph convolutional network with global feature enhancement for traffic flow prediction[J]. International Journal of Machine Learning and Cybernetics, 1-17.*

[4] *Bayati A, Nguyen K K, Cheriet M. (2020). Gaussian process regression ensemble model for network traffic prediction[J]. IEEE Access, 8: 176540-176554.*

[5] *Li D.(2020). Predicting short-term traffic flow in urban based on multivariate linear regression model[J]. Journal of Intelligent & Fuzzy Systems, 39(2): 1417-1427.*

[6] *Yu J, Feng Z, Ju L, et al. (2024). Traffic Flow Prediction Method Based on Improved Optimization Algorithm Mixed Kernel Extreme Learning Machine[C]//Proceedings of the International Conference on Algorithms, Software Engineering, and Network Security. 54-58.*

[7] *Huang X, Ma C, Zhao Y, et al. (2023). A hybrid model of neural network with VMD–CNN–GRU for traffic flow prediction[J]. International Journal of Modern Physics C, 34(12): 2350159.*

[8] *Yu Z, Yang J, Huang H H. (2024). Smoothing regression and impact measures for accidents of traffic flows[J]. Journal of applied statistics, 51(6): 1041-1056.*

[9] *Abbass K, Qasim M Z, Song H, et al. (2022). A review of the global climate change impacts, adaptation, and sustainable mitigation measures[J]. Environmental Science and Pollution Research, 29(28): 42539-42559.*

[10] *Arnell N W, Lowe J A, Challinor A J, et al. (2019). Global and regional impacts of climate change at different levels of global temperature increase[J]. Climatic Change, 155: 377-391.*

[11] *Obradovich N, Fowler J H. Climate change may alter human physical activity patterns[J]. (2017). Nature Human Behaviour, 1(5): 0097.*

[12] *Zhou X, Yu Z, Yuan L, et al. (2020). Measuring accessibility of healthcare facilities for populations with multiple transportation modes considering residential transportation mode choice[J]. ISPRS International Journal of Geo-Information, 9(6): 394.*

[13] *Johnstone C, Sulungu E D. (2023). Application of neural network in prediction of temperature: a review[J]. Neural computing and applications, 2021, 33(18): 11487-11498.*