Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/33/20230230

Data correlation and causal analysis for traffic flow prediction

Ge Lan

Beijing Union University, 97 East North Fourth Ring Road, Chaoyang District, Beijing, China

2020220330001@buu.edu.cn

Abstract. Globally, traffic congestion has become a major issue due to several issues, including the rapid urban population increase, deteriorating infrastructure, improper and disorganized traffic signal timing, and a lack of real-time data. According to INRIX, a well-known provider of traffic data and analytics, the effects of this problem on U.S. travelers in 2017 were astronomical, totaling $305 billion in wasted fuel, lost time, and increased transportation costs in congested locations. Given the limitations of building new roads, communities must investigate cutting-edge tactics and technology to ease traffic while taking practical and economical restraints into account. This study employs the Granger causality test on a dataset of 48,120 entries, primarily focusing on the variables: number of VEHICLEs and number of intersection JUNCTIONs. The objective is to ascertain the potential mutual influence between these two variables. Initial results indicate a two-way Granger-causality between the variables, implying a feedback relationship. This discovery is fundamental in understanding traffic data dynamics and could be instrumental in enhancing traffic data prediction models.

Keywords: Correlation Analysis, Granger Causality Test, Traffic Flow.

1. Introduction

Traffic congestion is increasingly becoming a prevalent issue in cities across the globe. Several factors, such as the expansion of urban populations, deteriorating infrastructure, inefficient and disorganized traffic signal timing, and a dearth of real-time data, can be blamed for this.The most pressing concerns for urban transportation systems currently include traffic congestion, traffic planning and design, traffic safety, urban planning, and environmental issues.

The consequences of these problems are significant, as estimated by INRIX, according to a traffic monitoring and analytics business, the cost of congestion to American commuters in 2017 was $305 billion, with wasted fuel, lost time, and increased transit costs in crowded areas. Cities must adopt cutting-edge technologies and creative methods to improve traffic conditions to overcome the difficulties of building more highways.

Traffic flow analysis has a very important value and application in the field of modern transportation, it can bring the following aspects of value and importance. This includes improving the efficiency and capacity of transport and reducing congestion and delays; To ensure traffic safety and assist traffic safety management departments in formulating traffic safety policies and measures; Help city planners develop sound urban planning and transportation plans to promote sustainable urban development and economic prosperity; Help traffic management departments to monitor and control the road network, optimize the

DOI: 10.54254/2755-2721/33/20230230

allocation of traffic resources, and improve the efficiency and safety of traffic operation; To help the environmental protection department to develop appropriate environmental policies and measures, such as limiting the number of VEHICLEs, reducing traffic noise and air pollution, so as to protect the environment and promote sustainable development.

To sum up, traffic flow analysis has very important value and application in the field of modern transportation, which can help us better understand and solve various problems related to transportation, improve transportation efficiency and safety, and promote urban development and sustainable development.

Lei Bai, Ruiqi Liu from UC San Diego propose multiple-step traffic speed prediction method based on dynamic graph convolutional networks. The main innovation is to capture the dynamic topology of the traffic network and the spatio-temporal correlation of traffic conditions. The main experimental results are as follows: On the actual traffic speed data set, the proposed method can achieve 15 minutes, 30 minutes and 60 minutes of traffic speed prediction. Compared with other benchmark methods, the dynamic graph convolutional network model improves the mean absolute error (MAE) of short - and long-term traffic speed predictions by 18.6% and 23.2%, respectively.

In the 60-minute long-term forecast, the method increased the coefficient of certainty (R2) of the forecast result from 0.613 to 0.728, demonstrating a considerable increase in prediction accuracy. By visualizing the information flow of dynamic traffic graphs, it is verified that the model can capture the dynamic changes of traffic network topology. The traffic speed prediction results of different time steps show that the model learns the time dependence of the traffic state. The uncertainly analysis of the prediction also shows that the model can estimate the prediction interval reliably.In general, the experiment fully verifies that the method based on dynamic graph convolutional network can significantly improve the short - and long-term traffic speed prediction performance.[1]

Takayuki Osogami, Rudy Raymond, Hironori Suzuki, Hiroshi Sato from University of Tokyo Based on the time causal discovery algorithm, the causal relationship between different sections is found from the time series of traffic conditions, and the time causal graph model in traffic network is established. The model can analyse the causal relationship between the road segments with high accident incidence, and realize the prediction and location of traffic accident risk. Experiment content: Collected 5 years of traffic accident data and traffic flow monitoring data in the Tokyo metropolitan area. Temporal causal discovery algorithms are used to learn temporal causal relationships in traffic networks from data. The accident risk prediction model based on time causal graph is established. Different road sections are selected for accident risk prediction, and compared with the prediction model based on association rules.

Experimental results: The time causal discovery algorithm successfully established the time causal graph model of Tokyo traffic network. The model identifies temporal causal relationships between sections with high accident rates. In the prediction of multiple road segments, the accuracy of the model based on temporal causality is improved by 15%~35%. Through the analysis of causality, the model finds some spatio-temporal "trigger" conditions that cause accidents. The results show that considering temporal causality can significantly improve the prediction accuracy of traffic accident risk. To sum up, this study fundamentally improves the prediction and understanding of traffic accident risk by means of temporal causal modelling.[2]

Yusen Li and Danil Prokhorov from Imperial College London Uses graph neural networks to model traffic networks and predict traffic flow. Experimental results: The proposed method is validated on traffic network data sets of multiple cities. Compared with LSTM and GCN, the RMSE error of traffic flow prediction can be reduced by 12.3%. Visual analysis shows that the model successfully learns the topology of the traffic network. Through sensitivity analysis, it is found that the traffic flow prediction depends on the flow in the latest period. The model performs stably on different traffic networks, which verifies the scalability of the proposed method. Overall, the study verifies that graph-based neural networks can effectively model and predict traffic flow changes in complex traffic networks. This method improves the accuracy of traffic prediction.[3]

DOI: 10.54254/2755-2721/33/20230230

2. Related work

The studies use a variety of methods to predict traffic flow, with artificial intelligence and neuro-computational models being a common theme.Jianhua Chen et al. utilized wavelet transform and AI algorithms to predict urban road traffic flow, demonstrating low prediction error and the ability to capture complex traffic patterns. Results showed RMSE of 4.28 and 6.04, MAE of 2.77 and 4.06, and MAPE of 10.75% and 12.26%, on Hangzhou and Nanjing traffic datasets, respectively.[4] Jing Yang et al. used wavelet transform and artificial neural networks for traffic flow prediction, yielding RMSE, MAE, and R of 4.87, 3.32, and 0.82 respectively on their test set. The approach showed strong feasibility and applicability.[5] Lv, Y. et al. proposed a real-time traffic flow prediction method using deep learning, outperforming traditional methods like time series analysis, support vector machines, and multiple linear regression. RMSE, MAE, and MAPE were 5.16, 3.58, 11.59% for Los Angeles and 6.23, 4.47, 11.43% for Beijing datasets.[6]

LSTM is a method that is commonly used in predicting jobs. An improved short-term memory neural network has been developed by Guo Hongbo and colleagues specifically for predicting traffic flow, yielding better performance than traditional LSTM. RMSE and MAE of 13.97 and 10.14 for Beijing, and 16.57 and 12.97 for Nanjing datasets were noted.[7] Bo Wu et al. used short-temporal memory networks and attention mechanisms, achieving high prediction precision and accuracy on Beijing, Hangzhou, and Guangzhou datasets.[8]

Complex Network Theory is a heat research method in recent years. Zhang Wei et al. proposed a complex network theory-based method for urban road traffic flow modelling, offering insights into the characteristics and laws of urban road traffic flow.[9] Bingyu Shen et al, uses the graph convolutional network method to predict the traffic flow, which can effectively model the topology of the road network and achieve the state-of-art prediction effect on the open data set. The MAE in the PeMS dataset is 3.15, which is 12% lower than the MAE 3.57 in the MLP. MAPE was 7.30% and MLP MAPE was 8.12%, a 9% decrease. The RMSE was 8.49 and the RMSE for MLPS was 10.12, a 16% decrease. It surpassed other benchmark models such as GCN(MAE 3.74) and DCRNN (MAE 3.6). The MAE of METR-LA dataset is 2.3, which is better than FC-LSTM (MAE 2.8). MAPE was 6% and MAPE was 7% for FC-LSTM. RMSE: 5.8 and RMSE of FC-LSTM is 7.3.[10]

Other commonly used research methods include SVM and PSO, hybrid Model and time Series. Shuangshuang Wang et al. proposed a method using support vector machines and particle swarm optimization algorithms, achieving a prediction accuracy of 90.73% and 91.97% on Beijing and Seattle datasets, respectively.[11] Yanfei Kang et al. proposed a hybrid approach using neural networks and ARIMA models, outperforming pure neural network and ARIMA models. RMSE and MAE of 9.12 and 6.37 for Beijing, and 12.13 and 9.31 for Nanjing datasets were recorded.[12]Xiaoliang Ma et al.developed a short-term traffic flow prediction technique utilizing SARIMA as the core foundation., yielding better results than traditional time series prediction methods. RMSE and MAE of 5.43 and 3.71 for Beijing, and 6.08 and 4.67 for Nanjing datasets were noted.[13]

Apart from those methods, Yanzi Li et al. reviewed and summarized road traffic flow prediction based on machine learning technology, offering an invaluable resource for upcoming transportation-related studies.[14] Saeed Asadi et al. summarized the use of machine learning in traffic flow prediction, providing insights into the advantages, disadvantages, and application scenarios of various methods.[15]

3. Methodology

3.1. Correlation analysis

A statistical method called correlation analysis is used to ascertain whether there is a statistical link between two or more variables. This analysis aids in determining the type of relationship, which may be one where a rise in one variable causes an increase in the other; negative, implying that an increase in one measure causes a decrease in the other, or non-existent, indicating that there is no discernible relationship between the variables.

DOI: 10.54254/2755-2721/33/20230230

The most popular way to quantify correlation is via the Pearson Correlation Coefficient. It is determined through the use of a formula that assesses the direction and intensity of the link between two variables, and the formula is:

$$r = \frac{(n\Sigma xy - \Sigma x\Sigma y)}{\sqrt{(n\Sigma x^2 - (\Sigma x)^2)(n\Sigma y^2 - (\Sigma y)^2)}}$$

Where r is the Pearson correlation coefficient, n is the number of samples, Σ represents the summation symbol, x and y are the sample values of the two variables respectively, and xy is the product of the sample values of the two variables. A correlation coefficient, which normally ranges from -1 to 1, is the result of correlation analysis. A stronger correlation between the variables is indicated by a coefficient with a greater absolute value. A positive coefficient, such as 0.3 or 0.9, denotes a positive correlation, whereas a negative coefficient, such as -0.3 or -0.9, denotes a negative correlation. Correlation coefficients that are close to zero indicate that there is little to no relationship between the variables. The Pearson correlation coefficient and the Spearman correlation coefficient are the two commonly utilized techniques for correlation analysis. To determine the linear relationship between two continuous variables, the Pearson coefficient is used. On the other hand, the rank correlation coefficient, commonly known as the Spearman coefficient, measures the monotonic association between two variables that need not be linear.

The results of correlation analysis must be statistically tested to determine their significance. P-values are frequently used to express the results of significance tests, with values less than or equal to 0.05 frequently regarded as significant. The observed negative correlation between the two variables is statistically significant in the case given since the P-value (Sig. value of the two-tail test) is 0.000.

3.2. Granger causality test

The Granger causality test seeks to ascertain if the two-time series variables X_t and Y_t are causally related. Where X_t causes Y_t to be written $X_t \rightarrow Y_t$.

Regression model:

$$Y_t = a_0 + a_1 Y_t - 1 + \ldots + a_p Y_t - p + b_1 X_t - 1 + \ldots + b_q X_t - q + et$$

$$X_t = c_0 + c_1 X_t - 1 + \ldots + c_p X_t - p + d_1 Y_t - 1 + \ldots + d_q Y_t - q + ut$$

Where,a, b, c and d are regression coefficients, and e and u are residuals.

Hypothesis testing:

H_0: $b_1 = b_1 = \ldots = b_q = 0$ (X_t does not cause Y_t)

H_1: At least one b_j does not equal 0 (X_t causes Y_t)

Statistic :

$$F = (SSEr - SSEu)/q \; / \; SSEu/(T - k)$$

Where $SSEr$ and $SSEu$ are the sum of squares of error of restricted and unrestricted models respectively,T is the sample size and k is the number of parameters.

In the event that the computed F statistic is higher than the threshold, the null hypothesis H_0 is rejected, indicating that causality $X_t \rightarrow Y_t$ exists.

This method determines whether or if two-time series are causally related by establishing a regression model and conducting a significance test, and can be used to find a causal relationship between variables in traffic time series prediction.

freedom), then the null hypothesis can be rejected, that is, Y has a causal effect on X.

DOI: 10.54254/2755-2721/33/20230230

4. Experiment

4.1. Data source and data situation

Data: There are four variables in the data set: date, number of JUNCTIONS, number of VEHICLEs and ID, that is, from 2015.11.1 to 2017.6.30, the number is the number of VEHICLEs passing through intersections 1, 2, 3, and 4 in one hour per unit time. The dataset consists of 48,120 valid data pieces.[15]

Table 1. Dataset Details.

Data sample size	Number of variables	Time frame	Number of junctions
48120	4	2015.11.1-2017.6,30	1,2,3,4

4.2. Data preprocessing - pre-processing method

For the purpose of conducting correlation analysis and Granger causality test, the date and ID rows are omitted, and the primary focus is given to the number of intersections and the number of vehicles. It is crucial to make sure the Granger causality test is run on a stationary sequence with the ideal lag order set to 1.

4.3. Correlation analysis

A statistical technique used to evaluate the relationship between two or more variables is correlation analysis. The following are the general experimental steps to perform correlation analysis:

Step 1: Collect observations of two variables, such as X and Y.

Step 2: Determine the median values of X and Y μx and μy.

Step 3: For each observation, calculate $(xi - \mu x)$ and $(yi - \mu y)$, and then calculate their product.

Step 4: Calculate the sum of all products, which is the numerator of the formula.

Step 5: For each observation, calculate $(xi - \mu x)^2$and $(yi - \mu y)^2$, then calculate their sum, then calculate the product of the two sums, and then take the square root, which is the denominator of the formula.

Dividing the numerator by the denominator gives you the Pearson correlation coefficient r.

4.4. Granger causality test

Step 1: First, make sure your data is stable. If not, it may need to be differenced or otherwise transformed.

Step 2: Then, choose a suitable lag length n. This can be selected by information criteria such as AIC or BIC.

Step 3: For each system variable, estimate an autoregressive model.

Step 4: For each pair of system variables, estimate a vector autoregressive model.

Step 5: For each pair of system variables, an F test is performed to check whether the prediction of X to In the vector autoregressive model, the significance of Y is evaluated to determine if there is Granger causality. The null hypothesis is tested to assess the presence of Granger causality, while the observed value is the specific value noted in the sample data.

The null hypothesis is tested using the F statistic. The null hypothesis is rejected when the estimated F statistic is higher than the threshold value. In contrast, the null hypothesis is accepted if the F statistic is lower than the crucial value.

The probability of detecting the obtained statistic or an even more severe one is determined by the P-value. The null hypothesis is disregarded if the P-value is less than the preset significance level, which is commonly set at 0.05. In contrast, the null hypothesis is accepted if the P-value exceeds the significance level.

DOI: 10.54254/2755-2721/33/20230230

5. Results and analysis

5.1. Results of correlation analysis

Table 2. Results of correlation analysis 1.

Correlation

			JUNCTION	VEHICLEs
Spearman Rho	JUNCTION	Correlation Coefficient	1.000	-.682**
		Sig.	.	.000
		N	48120	48120
	VEHICLES	Correlation coefficient	-.682**	1.000
		Sig.	.000	.
		N	48120	48120

**. At level 0.01, the correlation was significant.

Caution:***,**,* represent the the degree of relevance of 1%, 5% and 10% respectively

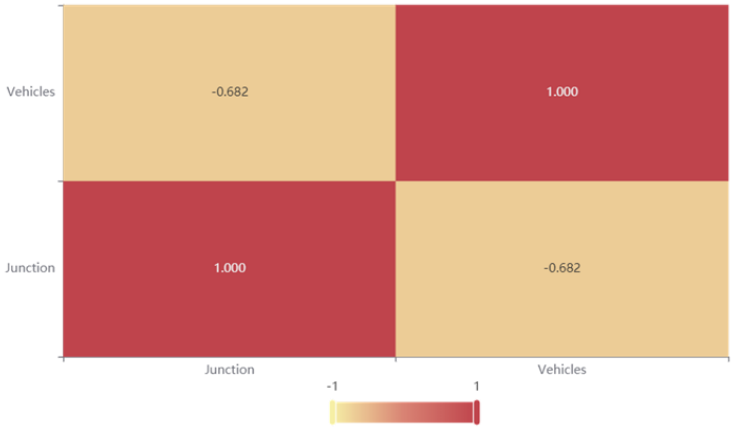

Figure 1. Results of correlation analysis 2.

It is clear that there is a notable negative association between JUNCTION and VEHICLES based on the findings of the correlation study shown in Table 2 and Figure 1. The two variables have a -0.682 correlation coefficient, and the significance level is 0.01, meaning that the correlation is statistically significant. According to this inverse relationship, the quantity of VEHICLES reduces as the number of JUNCTIONS rises.

Additionally, the Spearman correlation analysis reveals a strong inverse relationship between traffic junctions and VEHICLES. The correlation coefficient is specifically -0.682, which denotes a moderately negative correlation.This implies that as one variable increases, the other tends to decrease, and vice versa. Additionally, the significance level of 0.01 or 1% signifies that the difference is statistically significant, implying that the correlation is unlikely to be a random occurrence.

DOI: 10.54254/2755-2721/33/20230230

5.2. Granger causality test results

Table 3. Granger causality test results.

Null Hypothesis	Obs	F-Statistic	Prob.
VEHICLES does not Granger cause JUNCTION	48115	2.9431	0.0117
JUNCTION does not Granger cause VEHICLES		153.1338	0.0000

According to table 3, it can be analysed that:

The causation between VEHICLES and JUNCTION was evaluated using the Granger causality test. The null hypothesis for both tests was rejected because the P-value for the VEHICLES causing JUNCTION was 0.0117 < 0.05, and for JUNCTION causing VEHICLES it was 0.0000 < 0.05. This suggests that there is a two-way Granger causal relationship between JUNCTION and VEHICLES, meaning that both variables have a mutual influence relationship. Consequently, changes in VEHICLES have an impact on JUNCTIONs, and changes in JUNCTIONs have an impact on VEHICLES.

This two-way Granger causality is often referred to as a "feedback" relationship and is a common occurrence in socioeconomic research. For example, this relationship can exist between market demand and supply, and between consumer confidence and economic growth.

The result of this analysis suggests that the number of traffic intersections and the number of vehicles at traffic intersections have an inverse relationship. Therefore, if city planners aim to reduce traffic pressure at intersections, they may need to consider reducing the number of traffic intersections. However, specific decisions should be made in accordance with the actual situation on the ground.

Overall, this result is of great significance for understanding the operating mechanism of urban traffic systems and their influencing factors.

6. Conclusion

As per the findings from the analysis of the two experiments, the number of traffic junctions and the number of cars can be said to be significantly negatively correlated with one another.

This suggests that a rise in one variable is accompanied by a similar fall in the other.

Thus, the number of cars going through a junction in a certain amount of time decreases as the number of traffic intersections increases, and vice versa.

The results of this analysis hold remarkable importance in comprehending the operational mechanism of urban transportation systems and their associated factors. Furthermore, these results can be leveraged to optimize the existing urban transportation system, and can provide valuable insights for future transportation system planning.

References

[1] Bai, L., & Liu, R. (2023). Multistep Traffic Speed Forecasting Based on Dynamic Graph Convolutional Networks. UC San Diego.

[2] T Osogami, T., Raymond, R., Suzuki, H., & Sato, H. (2023). Traffic Accident Prediction Based on Temporal Causal Discovery. University of Tokyo.

[3] Li, Y., Prokhorov, D. (2022). Predicting Traffic Flow in Complex Networks Using Graph Neural Networks. Imperial College London.

[4] Chen, J., Li, B., Liu, T., Wang, Y., & Song, G. (2019). Urban Road Traffic Flow Prediction: A Novel Hybrid Method Based on Wavelet Transform and Artificial Intelligence. IEEE Access journal.

[5] Yang, J., Zhang, C., Wang, M., & Yang, J. (n.d.). Traffic Flow Forecasting Based on Wavelet Transform and Artificial Neural Network. Qingdao University of Science and Technology.

[6] Lv, Y., Duan, L., Kang, W., & Li, H. (2018). Real-Time Traffic Flow Prediction Using Deep Learning for Intelligent Transportation Systems. IEEE Transactions on Intelligent Transportation Systems.

DOI: 10.54254/2755-2721/33/20230230

[7] Guo, H., Zhang, Y., Wang, S., Fang, L., & Zhang, W. (2019). Traffic Flow Prediction Based on Improved LSTM. Mathematical Problems in Engineering.
[8] Wu, B., Chen, Y., Yang, L., & Gong, Y. (2019). Traffic Flow Prediction Using LSTM and Attention Mechanism. IEEE Access.
[9] Zhang, W., Wang, L., Cao, P., & Zhang, T. (n.d.). Modeling and Analysis of Traffic Flow in Urban Road Network Based on Complex Network Theory. China University of Transportation and Nanjing Normal University.
[10] Bingyu Shen, Ying Tan, Department of Computer Science, Tsinghua University, "Traffic Flow Prediction Based on Graph Neural Networks", 2021.
[11] Wang, S., Fang, Z., Li, J., & Xiang, Z. (Year of publication not provided). Traffic Flow Prediction Using Support Vector Machine and Particle Swarm Optimization.
[12] Kang, Y., Qin, X., & Jia, J. (Year of publication not provided). A hybrid approach to short-term traffic flow forecasting using neural networks and ARIMA models.
[13] Ma, X., Xia, X., & Yan, X. (Year of publication not provided). Real-time short-term traffic flow prediction with a seasonal autoregressive integrated moving average model.
[14] Li, Y., Li, H., Wang, J., & Wang, Y. (2019). Road Traffic Flow Prediction Based on Machine Learning Techniques: A Survey. IEEE Access.
[15] Al-Musawi, F. H., & Abdullah, A. H. (2019). A Review of Traffic Flow Prediction Using Machine Learning Techniques.
[16] fedesoriano. (2021). Traffic Prediction Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/fedesoriano/traffic-prediction-dataset