Differences Of Multiple Factors Affect Prediction of Traffic

Haoyang Li *

Bradshaw Christian High School, Sacramento, CA, 95829, United States

* Corresponding Author Email: dmoore81587@student.napavalley.edu

Abstract. Intelligent transportation system (ITS) helps to control traffic to protect pedestrian's life and decrease the time cost of transportation which can help reduce the potential risk of accident. The most significant thing in ITS is the prediction of traffic flow. There already have many ways to predict the traffic flows. However, most of those method doesn't realize that different factor has different effect and time cost. Time cost of a factor does not discuss in this paper. The purpose of this paper is comparing different factors that may affect traffic to find out which one affects traffic the most. Training models are used to compare the effect of those factors and relative error are used as accuracy. As a result of the study, time in a day affects traffic more in holiday, temperature, rain volume, snow volume, cloud cover, and the time in a day. This may mean that people more prefer to driving based on time not on weather.

Keywords: Traffic prediction; Machine learning; Traffic flows.

1. Introduction

The intelligent transportation system (ITS) is a system that monitors and manages traffic flow, reduces congestion, shows users the best routes, increases system productivity, and saves lives, money, and time through extensive communication, control, and the use of vehicle sensing and electronic technologies [1]. Today, the ITS is irreplaceable for people traveling. The most important thing in ITS is traffic flow prediction [2]. This helps to estimate time costs and select the best road. Multiple ways to predict traffic have been discovered in the past few years. Most of those are machine learning due to the huge development of artificial intelligence. At first, people used artificial neural networks (ANNs) to predict traffic flows [3]. Then there are more and more technologies from machine learning have been used in the prediction of traffic, such as temporal graph convolutional networks (T-GCN) which are coupled with the graph convolutional network (GCN) and gated recurrent unit (GRU) [4], FARIMA models [5], and decision tree models [6].

However, in most of those studies, as many factors as possible are put in the models. Although this may increase the accuracy of the prediction, some factors may not be worth it to cost that much time to get a little accuracy. In some cases, efficiency may be more important than accuracy. For example, control traffic signals based on the queue lengths or predicted traffic volumes to increase the efficiency of the road [7]. In this case, prediction needs to be estimated in a short time. Adding more factors may slow the prediction. In this paper, the target is to explore the difference of factors in natural situations affecting the prediction of traffic, which are weather and time.

2. Method

In the study, Linear regression and decision tree methods are used. Descriptions of those methods are in the following paragraphs.

2.1. Linear Regression

Linear regression is a common regression method used to show the linear relationship between a dependent variable and one or more independent variables. The relationship between the dependent variable and the independent variable is represented by a linear expression. [8]. Because linear regression is more accurate than random forest, MSE criteria, SVR, and MLP approaches, it is preferred [9].

2.2. Decision Tree

Developing prediction algorithms for a target variable and establishing classification systems based on several criteria are two typical applications of the decision tree technique. Using parts of a population that resemble branches, this technique generates an inverted tree with root, internal, and leaf nodes [10]. The decision tree is used because of its high accuracy and scalability [11].

2.3. Evaluate Formula

During the study, relative errors are used to evaluate the accuracy. It is calculated by the following formula. Here E_r represents relative error, T_p represents predicted traffic volume, and T_t represents true traffic volume. The relative error is a typical way to evaluate the accuracy.

$$E_r = \frac{|T_p - T_t|}{T_t} \times 100\% \quad (1)$$

3. Experiment

3.1. Dataset Introduction

The datasets used in the paper are collected from a highway in the US by the local Department of Transportation. This is collected from 2012 to 2018 including over 48000 rows [12]. Feature traffic volume shows the counts of vehicles that pass a point in the highway from east to west. Other features that describe the weather and date shown in Table 1 are holiday, temperature, rain volume, snow volume, cloud cover, a short text of the weather, a longer text of the weather, and the time of local time.

Table 1. Datasets Features

Parameter description	Unit	Abbreviation
Is a US national Holiday or a regional Holiday		holiday
temperature	Kelvin(K)	temp
Rain volume in the hour	millimeter(mm)	rain_1h
snow volume in the hour	millimeter(mm)	snow_1h
the percentage of cloud cover		clouds_all
a short textual describe the current weather		weather_main
a longer textual describe the current weather		weather_description
the time of the data collected in local CST time		date_time
the hourly reported westbound traffic volume		traffic_volume

To make dataset features fit the model, date and time are separated into hours, day of week, and month. The month is an integer from 1 to 12, and the day of the week is an integer from 1 to 7. Then weather main and weather description are dropped. Table 2 shows the features of the final dataset.

Table 2. Datasets Feature After the Process

Parameter description	Unit	Abbreviation
Is a US national Holiday or a regional Holiday		holiday
temperature	Kelvin(K)	temp
rain occurred in the hour	millimeter(mm)	rain_1h
snow occurred in the hour	millimeter(mm)	snow_1h
the percentage of cloud cover		clouds_all
hour in a day		hour
day of week		day of week
month		month
the hourly reported westbound traffic volume		traffic_volume

3.2. Result Prediction

Before the study, by combining the experience of life, time is expected to affect traffic the most. Most people have a job which means they have a fixed schedule. The classical working day is 7-8 a.m. to 5-6 p.m., Monday to Friday [13]. This shows that the hours before 8 a.m. and after 5 p.m. should have the highest traffic caused by transportation for work. Other factors such as weather and holidays don't affect most people. Most weather does not affect people who drive, and holidays only affect some people who would like to travel.

3.3. Experiment Process

In the study, accuracy is used to estimate the effect of factors. The higher the accuracy, the higher the effect of those factors. First, features are dropped one by one to get new datasets. Those datasets are used to train and compare with original datasets. Fig. 3 shows the code for creating new datasets. Each dataset is trained 5 times for each algorithm, which means a total of 10 times. Additionally, the model is trained using 90% of the data and then tested using the remaining 10%.

3.4. Result

The average relative error for each dataset is shown in Table 3. The feature rain in the hour and snow in the hour are removed in no rain and snow datasets, feature hour in a day is removed in no time datasets, and feature day of week and month are removed in no date datasets. The data in Table 3 is shown in the graph in Fig.1

Table 3. Running Results

data type	error in decision tree	error in linear regression	difference from control group
control group	19.17%	133.33%	0.00%
no holiday	19.98%	133.61%	1.09%
no temperature	25.29%	133.50%	6.29%
no rain and snow	20.13%	133.27%	0.90%
no cloud	19.71%	131.69%	-1.10%
no time	106.89%	165.89%	120.28%
no date	26.20%	131.69%	5.39%

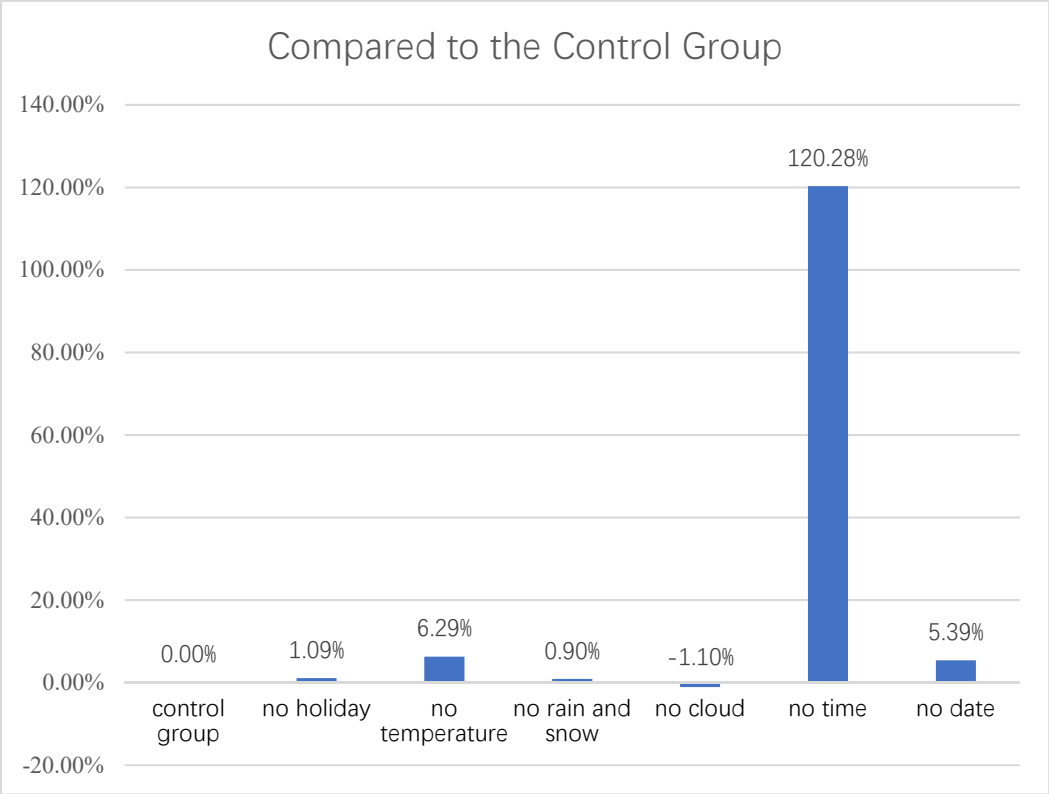

Fig. 1 Running Result

4. Discussion

From the result of the experiment, it's obvious that datasets without time have the highest relative error which is 106.89% when using a decision tree and 165.89% when using linear regression. The relative error is 5 times higher when using datasets that drop the feature time in the decision tree and 30% higher in linear regression. Thus, time could be said to affect traffic the most. However, this data set was collected from a highway in the US, which means this may not fit every country and every road. Also, there might be other factors that affect traffic. Thus, the conclusion of this paper may not be accurate.

Through the entire study, we conclude that time affects traffic the most. This can help people make a simple prediction of traffic quickly, so they can make a better traveling plan. Moreover, this conclusion can also be used to help engineers improve the algorithm for the prediction of traffic. Also, this can help model trainers select and add weight for each factor. Let them more focus on time not other factors to make the algorithm better and faster. This may help with driverless technology in the future.

References

[1] Singh, B., & Gupta, A. (2015). Recent trends in intelligent transportation systems: a review. *Journal of Transport Literature*, *9*(2), 30–34. https://doi.org/10.1590/2238-1031.jtl.v9n2a6

[2] Da Zhang. Combining weather condition data to predict traffic flow: a GRU-based deep learning approach. 27 March 2018. 27 August 2023. https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/iet-its.2017.0313

[3] Smith L.B. Demetsky M.J.: 'Short-term traffic flow prediction: neural network approach', Transp. Res. Rec., 1994, 1453, pp. 98–104

[4] L. Zhao et al., "T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction," in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 9, pp. 3848-3858, Sept. 2020, doi: 10.1109/TITS.2019.2935152.

[5] Yantai Shu, Zhigang Jin, Lianfang Zhang, Lei Wang and O. W. W. Yang, "Traffic prediction using FARIMA models," 1999 IEEE International Conference on Communications (Cat. No. 99CH36311), Vancouver, BC, 1999, pp. 891-895 vol.2, doi: 10.1109/ICC.1999.765402.

[6] Alajali, W.; Zhou, W.; Wen, S.; Wang, Y. Intersection Traffic Prediction Using Decision Tree Models. Symmetry 2018, 10, 386. https://doi.org/10.3390/sym10090386

[7] Wang, J., et al. "Short-Term Traffic State Prediction from Latent Structures: Accuracy vs. Efficiency." Transportation Research Part C: Emerging Technologies, Pergamon, 24 Dec. 2019, www.sciencedirect.com/science/article/abs/pii/S0968090X19308009.

[8] Kim, S.-J.; Bae, S.-J.; Jang, M.-W. Linear Regression Machine Learning Algorithms for Estimating Reference Evapotranspiration Using Limited Climate Data. *Sustainability* 2022, *14*, 11674. https://doi.org/10.3390/su141811674

[9] Artin, Javad, et al. "Presentation of a Novel Method for Prediction of Traffic with Climate Condition Based on Ensemble Learning of Neural Architecture Search (NAS) and Linear Regression." Complexity, Hindawi, 31 Aug. 2021, www.hindawi.com/journals/complexity/2021/8500572/.

[10] Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

[11] Alajali, Walaa, Wei Zhou, Sheng Wen, and Yu Wang. 2018. "Intersection Traffic Prediction Using Decision Tree Models" *Symmetry* 10, no. 9: 386. https://doi.org/10.3390/sym10090386

[12] Ansh Tanwar. (Aug.2023). Interstate Traffic Dataset (US), Retrieved Sep.2023 from https://www.kaggle.com/datasets/anshtanwar/metro-interstate-traffic-volume.

[13] Giovanni Costa. Shift Work and Health: Current Problems and Preventive Actions. 17 July 2013. 27 August 2023. https://www.sciencedirect.com/science/article/pii/S2093791110120034#bb0010