IJIE

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

The International Journal of Integrated Engineering

Traffic Flow Prediction using SUMO Application with K-Nearest Neighbor (KNN) Method

Fikri Aditya[1], Surya Michrandi Nasution[1*], Agus Virgono[1]

[1]Computer Engineering, Faculty of Electrical Engineering,
Telkom University, Bandung, 40257, INDONESIA

*Corresponding author

DOI: https://doi.org/10.30880/ijie.2020.12.07.011
Received 29 February 2020; Accepted 9 August 2020; Available online 30 August 2020

Abstract: In Indonesia, the density of traffic flow occurs at the time of leaving and returning to work, long holidays or national holidays such as the end of the year (New Year). This annual routine activity is mostly carried out especially in big cities in Indonesia such as Bandung. Because Bandung is a city that has a lot of tourism, Bandung is therefore always the center of visitors to enjoy weekends or long holidays. So from this problem, we want to create a traffic prediction application that can help to solve congestion problems that have become an annual routine. The several types of vehicles used in the prediction are private cars, motorcycles, taxis, public transportation, large buses, mini buses, and mini trucks. Research conducted using the K-Nearest Neighbor method is a prediction of short-term traffic flow on Jl. Riau Bandung. The input used in making predictions is historical data on the number of vehicles going on Jl. Riau Bandung. The output generated from the use of the K-Nearest Neighbor method is the level of the jam class that runs on Jl. Riau Bandung in 2018 used a simulation on the SUMO (Simulation of Urban Mobility) application. The resulting performance of KNN with k = 3 has an accuracy of 99.21%, k = 5 has an accuracy of 99.60%, and k = 7 has an accuracy rate of 99.21% on 90% training data and 10% testing data.

Keywords: Prediction, Traffic, K-Nearest Neighbor (KNN), Simulation of Urban Mobility (SUMO).

1. Introduction

People always want to do urbanization, where people who live in small areas will move to big cities in the hope of increasing financial income. The process of modern urbanization is accelerating. With rapid urban population growth, more and more urban vehicles are being carried out. Urban roads are complicated, and urban traffic problems are becoming increasingly serious. And road traffic congestion [1] - [2] is one of them. In big cities, when a traffic jam occurs, if it is not handled in a timely manner, it will cause more and more crowded areas and can even cause traffic paralysis.

Just like other big cities, Bandung is one of the cities in Indonesia with high population growth. In 2014, the city of Bandung was ranked seventh as the most congested city in Indonesia with a congestion rate of 14.3 km per hour and a VC ratio of 0.85. This is after the Indonesian Ministry of Transportation released the list of cities with the densest traffic in Indonesia [3] and according to the findings of the Inrix research Institute, the level of traffic congestion in cities in the world has increased. This increase in traffic jams also occurred in the cities in Indonesia surveyed by Inrix throughout 2017, and Bandung occupies the second position of the most congested city in Indonesia after Jakarta [4] and that means Bandung has experienced a very rapid increase compared to 2014. According to data from the Transportation Office Bandung City in 2018, there were 1,251,080 two-wheeled vehicles, and 536,973 four-wheeled vehicles in the city of Bandung. This number continues to increase 11% per year with private vehicles dominated by 98% and 2% public transportation [5] and if there is no real innovation to overcome this, then within the next 3 years we will have difficulty leaving the house, because just want to come out already jammed.

*Corresponding author: michrandi@telkomuniversity.ac.id

For now the Bandung government in an effort to overcome the problem of congestion is only thinking about longterm solutions and spending a considerable amount of money, such as the construction of inner-city toll roads [6]. Yet according to Chairman of the Indonesian Transportation Society Sony Sulaksono, inner-city toll roads are only a momentary solution for a worse future. Tolls in the city, for example in Jakarta, Bogor, Purbaleunyi Toll Road are toll roads that are not well integrated. The density of the vehicle continued to surpass the non-toll sections after the toll gate, all easily entering the City. Inside the toll road is no problem, but exiting the toll road is a new problem.

The first thing to overcome traffic jams is to prevent them from happening. Therefore, to overcome this problem a prediction of traffic flow is made using the SUMO application which aims to consider the similarity or similarity of traffic flow patterns from existing historical data using the K-Nearest Neighbor (KNN) method. With this prediction, it is expected to be able to help form an effective traffic jam forecast conducive to the preparation of targeted preventative and early warning measures. Most of these methods are prediction and analysis of traffic flow parameters. Based on traffic conditions data from the Department of Transportation related to different traffic conditions that have accumulated in the previous traffic platforms.

2. Theory

2.1 Street Characteristic

Riau Street Bandung is one of the central roads in Bandung where this road is a two-way street back and forth. Riau Street Bandung is one of the busiest streets in Bandung, which is crowded by vehicles because around this road there are hotels, cafes, schools, food restaurants and offices.

2.2 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is used as a basic algorithm to identify traffic profiles. KNN is a non-parametric pattern recognition technique commonly used for classification and regression purposes. If a given object is not labeled, the algorithm searches for the same object or neighbor from the search space and provides a label for the unlabeled object based on the nature of the nearest neighbor. The same concept can also be applied to the sequence of observations, for example, measurement of current levels. The algorithm identifies the K sequence from past data that is most similar to the same pattern that is being examined. The combination of the closest values corresponds to the time step in which an expected estimate will be made for the expected future value. KNN-based prediction refers to that the pattern of observation sequences repeated over time. Therefore, if the previous pattern can be identified to be similar to the current pattern, then the next value from the previous sequence can be used to predict future values.

Measurement of closeness between data can be measured based on the distance between the data with one another. The closer the distance between a data, the greater the similarity between the two data and vice versa. There are various ways to measure the closeness between data by using Euclidean distance. For the formula of Euclidean distance, they are as follows:

$$D = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} \quad (1)$$

Where :

X_1 =Volume of the vehicle in training data

X_2= Volume of the vehicle in testing data

Y_1 = V/C ratio in training data

Y_2 = V/C ratio in testing data

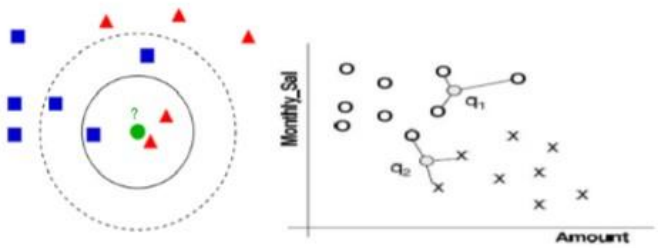

Fig. 1 - K-Nearest Neighbor

In Fig. 1, let's say K = 3, so that it can be seen from the 3 closest neighbors, the training data with the red triangle characteristic has the most number. Therefore, test data (green round) can be classified into red triangle training data.

The K value depends on the trial data needed. In general, greater K values have better accuracy, but make the boundaries between each classification less clear.

2.2 Simulation of Urban Mobility (SUMO)

Simulation of Urban MObility (SUMO) is a series of open, microscopic and sustainable traffic simulations designed to handle large-scale road networks [7]. SUMO can model intermodal traffic systems including public transport vehicles and pedestrians. Many tools can be used in SUMO applications, such as route search, visualization, traffic flow modeling and gas vehicle emission calculations [8]. SUMO can be improved with special models to control simulations from remote control. SUMO can work by creating its own traffic manually or cam import maps from Open Street Maps (OSM).

3. Methodology

In this study a prediction of congestion grade level using the K-Nearest Neighbor method is then simulated using the SUMO application on Riau Bandung street. Right after we collected the dataset, it will be preprocessed and classify using KNN. Its result will be used as performance evaluation, and it will simulate the traffic using SUMO. This methods is shown in Fig. 2.

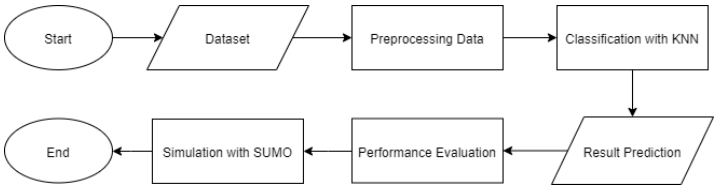

Fig. 2 – design system

3.1 Dataset

The dataset used is historical data on the number of vehicles flowing on the Bandung Riau road which was obtained by the Department of Transportation (DISHUB) of the City of Bandung in November to December 2018. The sampling period taken is every 15 minutes.

Table 1 - Sample dataset

Time	Week 1	Week 2	Week 3	…	Week 8
05.00 – 05.15	193	196	189	…	371
05.15 – 05.30	426	400	430	…	658
05.00 – 05.15	685	695	677	…	673
05.00 – 05.15	987	985	990	…	994
…	…	…	…	…	…
20.45-21.00	903	932	911	…	971

In Table 1, shows that the experimental data used are Tuesday starting at 05.00 until 21.00 with a sampling period every 15 minutes, and has historical data on the number of vehicles from week 1 to week 8 that will be used as input data. Of course we can choose one working day from Monday to Friday.

3.2 Preprocessing

Preprocessing is the initial stage of training in the data used. In this study, the following are the data preprocessing steps.

a. Number of Vehicles

Data on the number of vehicles is the most influential data in this study, because from historical data the number of vehicles will be used as training data before predictions are made.

b. Normalization

Data normalization aims to accelerate the learning stage in the system and prevent attributes with a broad range of values [9]. Because it changes data in a range [0,1] which aims not to use a lot of memory in the process of learning data. This case used normalization using the min-max method.

$$Normalization = \frac{x - Min\,(y)}{Max(y) - Min\,(y)} \quad (2)$$

Where x is the attribute data that will be normalized to a value in the range [0,1], then Min (y) is the minimum value of the whole data, and Max (y) is the maximum value of the entire attribute data respectively.

c. Data Partition

Data in the partition to determine training data and test data in accordance with scenarios with the aim of knowing the level of accuracy according to the method used. The following is a data sorting table divided into two namely training data and testing data.

Table 2 - Data partition scenario

No.	Training	Testing	Accuracy		
			K = 3	K = 5	K = 7
1	50%	50%	%	%	%
2	60%	40%	%	%	%
3	70%	30%	%	%	%
4	80%	20%	%	%	%
5	90%	10%	%	%	%

3.3 Classification with KNN

Normalized data and partitions are then predicted by the classification process using the KNN by finding the nearest neighbors of the data. In this study, a comparative effect of the values of k = 3, k = 5, and k = 7 on the performance of the research system conducted. The k value used is an odd number to avoid the same amount of data when determining results. To determine the closest distance to the value of K, you can search for the Euclidean Distance formula.

3.4 Performance Evaluation

Performance evaluation is used to measure the performance of the prediction system used. In performance evaluation, Confusion matrix is a method used to calculate the performance of a system built. Evaluation using Confusion matrix can help to get the results of the accuracy value of the system that has been built. Accuracy is the level of closeness between the predicted value and the actual value.

Table 3 - Confusion matrix

Actual	Predict	
	True	False
True	TP	FP
False	FN	TN

From the Confusion matrix above, accuracy and other performance evaluation obtained by using equation (3). By using this equation, we will need several value called TP, TN, FP, and FN respectively. Accuracy refers to the level of agreement between actual and predictive measurements. Accuracy the higher the accuracy of a system, the better the system.

$$accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \times 100\% \quad (3)$$

where :

TP (True Positive) : The data is predicted to be "true" and in the dataset class the value is "true"
FP (False Positive) : The data is predicted to be "wrong" and in the dataset class the value is "true"
TN (True Negative) : The data is predicted to be "true" and in the dataset class the value is "false"
FN (False Negative) : The data is predicted to be "true" and in the dataset class the value is "false"

4. Result and Discussion

This research was conducted by several dataset test scenarios with each result that has passed preprocessing data normalization and using the KNN method with k = 3, k = 5, and k = 7.

Table 4 - Accuracy performance results

No.	Training	Testing	Accuracy		
			K = 3	K = 5	K = 7
1	50%	50%	98.98%	98.35%	97.89%
2	60%	40%	98.92%	98.43%	98.04%
3	70%	30%	99.08%	98.56%	98.43%
4	80%	20%	99.02%	99.02%	98.36%
5	90%	10%	99.21%	99.60%	99.21%

From Table.4, the calculation of accuracy with k = 3, k = 5, and k = 7 on 5 partitions according to the scenario by passing the preprocessing stage of data normalization using the min-max method then proceed using the classification with the KNN method with the highest accuracy results of 99.60 % at K = 5 with 90% training partitions and 10% testing. The lowest accuracy is 97.89% at K = 7 with 50% training partition and 50% testing. It can be said that the higher the training data partition and the smaller K value, the higher the accuracy value obtained, and vice versa when the training data partition is lower and the distance K value is greater, the accuracy value will be lower as it gets closer high error value.

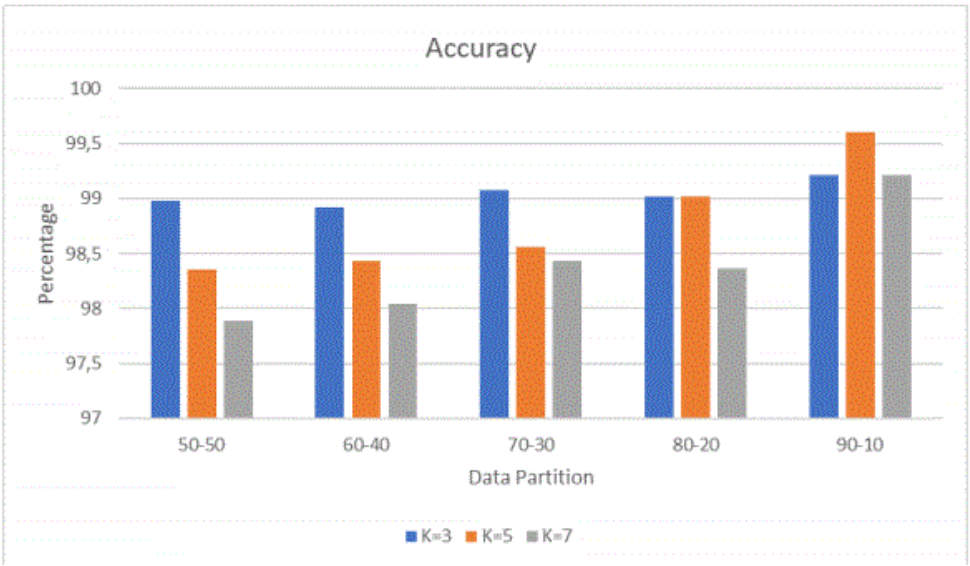

Fig. 3 - accuracy results using KNN

The predicted results of models and real traffic flow, it is observed that the predicted traffic flow has similar traffic patterns with the real traffic flow and the prediction value of the proposed KNN method is almost coincided with the measured data, especially in morning and evening peak hours.

After the prediction results are obtained, the vehicle is simulated in the SUMO application that has been determined by the map object and its route according to the number of vehicles generated from the predicted value. In Fig. 4 (a) shows the result of level of service prediction (Level A) and Fig. 4 (b) shows the result of Level D.

a b

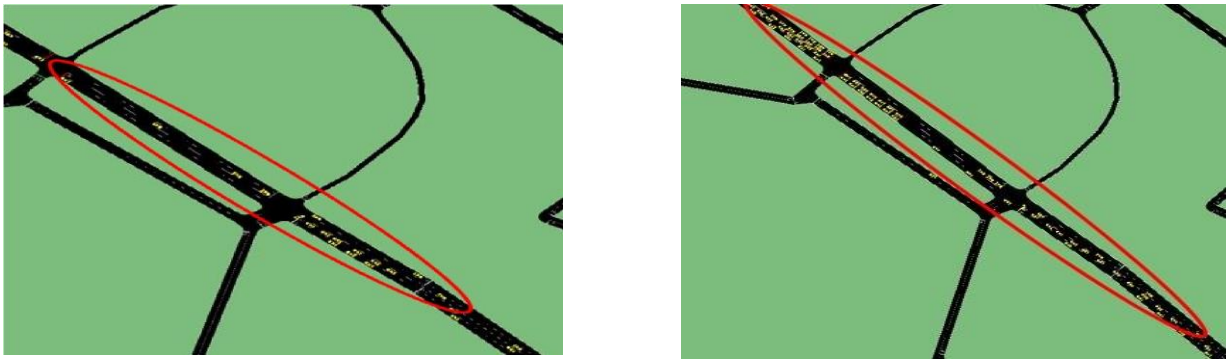

Fig. 4 - (a) predict level of service A; (b) predict level of service D

5. Conclusion

Based on the research results described in this paper, there are several conclusions including the distribution of historical data on vehicle traffic flow on Jalan Riau Bandung, can be predicted short-term using the K-Nearest Neighbor algorithm. The performance generated from the K-Nearest Neighbor algorithm with an accuracy value of k = 3 is 99.21%, k = 5 has an accuracy of 99.60%, and k = 7 has an accuracy rate of 99.21%. Modelling the results of predictions using the SUMO application with objects that have been determined on the map, then determining the route for vehicle simulation from the prediction results using the K-Nearest Neighbor method on Riau Bandung road.

Acknowledgment

The authors would like to thank to Telkom University especially Electrical Engineering Faculty, for giving the authors opportunities to conducts these research and let the authors finished it without any difficulties.

References

[1] ILhamnoor. (2014). Bandung, Kota Termacet Ketujuh Se-Indonesia. Available Online: https://infobandung.co.id/bandung-kota-termacet-ketujuh-se-indonesia [Accessed 7 February 2019]

[2] Satria Widianto. (2018). Sering Macet Parah, Ganjil Genap akan Diterapkan di Bandung. Available Online: https://www.pikiran-rakyat.com/bandung-raya/2018/09/26/sering-macet-parah-ganjil-genap-akanditerapkan-di-bandung-430694 [Accessed 7 February 2019]

[3] F. G. Habtemichael & M. Cetin. (2016). Short-term traffic flow rate forecasting based on identifying similar traffic patterns. Transportation Research Part C: Emerging Technologies, 66, 61-78

[4] B. W. Taylor III. (2006). Introduction to Management Science, Ninth Edition. Virginia Polytechnic Institute and State University: Prentice Hall

[5] Arief. (2013). Teknik Peramalan. Available Online: http://informatika.web.id/teknikperamalan.htm. [Accessed February 22, 2019]

[6] Abdul Muhaemin. (2018). Atasi Kemacetan, Pemkot Bandung Siapkan Sejumlah Rencana. Available Online: https://www.pikiran-rakyat.com/bandung-raya/2018/10/02/atasi-kemacetan-pemkot-bandung-siapkansejumlah-rencana-430961 [Accessed 9 April 2019]

[7] Daniel Krajzewicz, Georg Hertkorn & Peter Wagner. (2002). SUMO (Simulation of Urban Mobility) an opensource traffic simulation. Institute of Transportation Systems German Aerospace Centre

[8] Michael Behrisch, Laura Bieker, Jakob Erdmann, & Daniel Krajzewicz. (2011). SUMO – Simulation of Urban Mobility. Institute of Transportation Systems German Aerospace Centre

[9] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier