Feature Engineering Algorithms for Traffic Dataset

Akibu Mahmoud Abdullah*[1], Raja Sher Afgun Usmani[2], Thulasyammal Ramiah Pillai[3],
Ibrahim Abaker Targio Hashem[4], Mohsen Marjani[5]
School of Computer Science and Engineering, Taylor's University,[1,2,3,5]
Selangor, Malaysia
College of Computing and Informatics, Department of Computer Science[4],
University of Sharjah, 27272 Sharjah, UAE

***Abstract*—As a result of an increase in the human population globally, traffic congestion in the urban area is becoming worse, which leads to time-consuming, waste of fuel, and, most importantly, the emission of pollutants. Therefore, there is a need to monitor and estimate traffic density. The emergence of an automatic traffic management system allows us to record and monitor motor vehicles' movement in a road segment. One of the challenges researchers face is when the historical traffic data is given as an annual average that contains incomplete data. The annual average daily traffic (AADT) is an average number of traffic volumes at the roadway segment in a specific location over a year. An example of AADT data is the one given by Road Traffic Volume Malaysia (RTVM), and this data is incomplete. The RTVM provides an average of daily traffic data and one peak hour. The recorded traffic data is for sixteen hours, and the only hourly data given is one hour, from 8.00 am to 9.00 am. Hence there is a need to estimate hourly traffic volume for the remaining hours. Feature engineering can be used to overcome the issue of incomplete data. This paper proposed feature engineering algorithms that can efficiently estimate hourly traffic volume and generate features from the existing dataset for all traffic census stations in Malaysia using queuing theory. The proposed feature engineering algorithms were able to estimate the hourly traffic volume and generate features for three years in Jalan Kepong census station, Kuala Lumpur, Malaysia. The algorithms were evaluated using the Random Forest model and Decision Tree Models. The result shows that our feature engineering algorithms improve machine learning algorithms' performance except for the prediction of NO_2 using Random Forest, which shows the highest MAE, MSE, and RMSE when traffic data was included for prediction. The algorithm is applied in one of the traffic census stations in Kuala Lumpur, and it can be used for the other stations in Malaysia. Additionally, the algorithm can also be used for any annual average daily traffic data if it includes average hourly data.**

Keywords*—Feature engineering algorithm; queuing theory; Road Traffic Volume Malaysia (RTVM); machine learning algorithms

I. Introduction

As a result of an increase in the human population globally, traffic congestion in the urban area is becoming worse, which leads to time-consuming, waste of fuel, and, most importantly, the emission of pollutants. Therefore, there is a need to monitor and estimate traffic density. This reason results in the emergence of an automatic traffic management system for recording and monitoring the hourly and daily movement of motor vehicles. Several studies reported that motor vehicles are primary sources of air pollution in the urban area worldwide [1]. The concentration and increase of air pollution depend on the increase of traffic volume, speed of the vehicle, type of vehicle and many more factors. Researchers are looking for creative solutions like smart cities and GIS systems to avoid traffic congestion and volume [2, 3]. A study conducted reveals that traffic volume has a significant impact on PM_{10}, NO_x, NO, and NO_2 concentrations [4]. [5]'s study shows that the increase of the vehicle increases the concentrations of air pollution during peak hours in the morning and evening. Traffic volume, traffic congestion, and low speed increase level of PM and NO_x emissions [6]. A study in Kula Lumpur shows that air pollution concentration strongly depends on traffic volume, waiting time on the road, speed of the vehicle, and fuel consumption [7]. Speed of the vehicle, composition, traffic volume, intensity, and acceleration influenced the concentration of air pollution [8].

Number	Year	Time	HTV	Car	Van	MIorry	HIorry	Bus	Motorcycl	AWT	Speed
1	2014	0700-0800	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2	2014	0800-0900	9415	67%	8.90%	2%	0.20%	0.70%	21.20%	N/A	N/A
3	2014	0900-1000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
4	2014	1000-1100	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
5	2014	1100-1200	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
6	2014	1200-1300	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
7	2014	1300-1400	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
8	2014	1400-1500	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
9	2014	1500-1600	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
10	2014	1600-1700	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
11	2014	1700-1800	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
12	2014	1800-1900	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
13	2014	1900-2000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
14	2014	2000-2100	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
15	2014	2100-2200	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
16	2014	2200-2300	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
TOTAL			133,180	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Fig. 1. Summary of Daily and Hourly Traffic Volume.

One of the challenge researchers are facing is when the recorded traffic data is given as an annual average that contains incomplete data- and most of these researchers are conducting multidisciplinary studies- [9]. The annual average daily traffic (AADT) is an average number of traffic volumes at the roadway segment in a specific location over a year. AADT data are collected using surveillance cameras to count and monitor passing vehicles on a 24-hour, 16-hour, 5-hour, or 1-hour basis. These data are used mostly in road transport studies, such as estimation of fuel consumption, roadway planning, emission prediction, traffic operation, travel behavior, accident predictions, and many more [10]. An example of AADT data is the one given by Road Traffic Volume Malaysia (RTVM), and this data is incomplete. The RTVM provides an average of daily traffic data and one peak hour. The recorded traffic data is for sixteen hours, and the only hourly data given is one hour, from 8.00 am to 9.00 am, as shown in Fig. 1. The total daily traffic volume and peak hour (hourly traffic volume from 8.00

to 9.00 am) were highlighted with red color in the Figure. The highlighted blue color indicates the volume of type of the vehicle. The not available (N/A) in the Figure shows that the remaining hourly traffic volume and volume of the type of the vehicles were missing. There is a need to estimate the hourly traffic for the remaining hours. In this study, feature engineering was applied to overcome the issue of incomplete hourly traffic volume.

Feature engineering is one of the most challenging and significant tasks in data science. Extracting and generating new variable from the existing dataset is a difficult task, and also consume time, and effort to process variable in dataset before applying them in the model. Feature engineering is the process of extracting and generating new features or variables from the existing dataset which helps in improving the performance of Machine Learning Algorithms. It also helps to understand the data deeply and gives more valuable insights. Data scientist spend more than 80% of their time on cleaning the dataset [11].

The structure of the paper is presented as follows, Section II presents the related works, Section III discusses the methodology, and it is divided into two sections; namely, Section III-A presented the data and how it was collected, and Section III-B discusses the proposed feature engineering algorithms. In Section IV, the results were presented, it has two sections, Section IV-A is the feature engineering algorithms result and Section IV-B is the prediction of traffic emissions with and without traffic dataset. Discussion was presented in Section V. Lastly, the conclusion is discussed in Section VI.

II. Related Work

Traffic volume is one of the important variable, which contribute for increasing air pollution level produced by automobiles. Many studies were conducted to estimated hourly traffic volume, for example [12] applied extreme gradient boosting tree (XGBoost) and graph theory to estimate hourly traffic volume at location without traffic sensor in Utah United States of America (USA). The developed model was able to estimate hourly traffic volume. Study of [13] propose deep learning algorithm and image processing method to estimate traffic volume, vehicle type, and vehicle speed using recorded traffic video. The model was found good with 90% of accuracy for traffic volume estimation. Estimation of hourly traffic volume was conducted using Artificial Neural Network (ANN) [14]. The applied ANN model was able to estimated hourly traffic volume.

Time spent on the road and speed vehicles were responsible for the variability and trend of air pollution. Several studies were conducted to estimate vehicle speed and time spent on the road. These studies include [15], [16], [17], [18], [19], [20], [21], [22], [23], and [24]. These studies can be divided into three. Firstly, most of the studies have speed parameters in their data, so they developed models to estimate the vehicle speed in a location that they do not have traffic sensor stations. Some researchers proposed an algorithm to estimate the dataset's missing values, while others focus on estimating the speed to evaluate their models using the historical dataset. The second category is having a recorded video of the moving automobiles on the road, so they developed methods to estimate the vehicle's speed. Lastly, some studies installed sensor devices on the road to calculate and estimate the vehicles' speed.

In general, all these studies used four types of traffic datasets, namely, sensor device data, video-based data, image-based data, and vehicle data. Fig. 2 presents the four types of datasets used for the estimation of vehicle speed and time spent on the road. All of the presented studies none of them estimate or generate vehicle speed and time spent using AADT dataset. To the best of our knowledge, we could not find a study that generates the cars' speed and time spent on the road using AADT dataset.

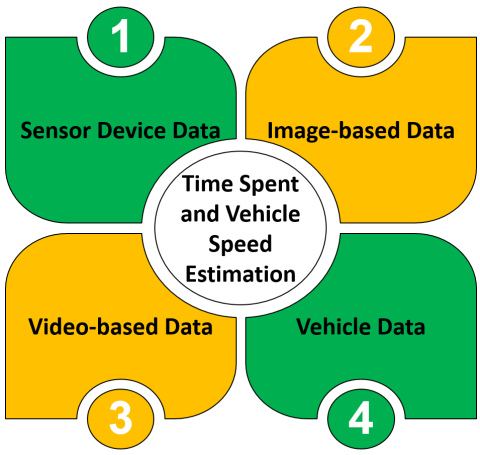

Fig. 2. Types of Dataset Used for Estimation of Vehicle Speed and Time Spent on the Road.

The RTVM data have been used by many studies in Malaysia. Table V presents the researches were conducted in Malaysia using the RTVM dataset. These studies mostly used the data to explore the Level of Service (LOS), the number of registered vehicles, and traffic density. There is a lack of study using hourly traffic data due to unavailability and incomplete hourly traffic volume dataset. In this paper, we proposed feature engineering algorithms which can efficiently estimate hourly traffic volume and generate features from existing traffic dataset (RTVM dataset). Queuing model is proposed to generate vehicle speed and time spent on the road features.

III. Methodology

A. Data

There is a total number of 554 traffic monitoring stations all over Malaysia. The traffic was recorded hourly for the state and federal roads in Malaysia by the State Public Works Department (JKR Negeri) and organized by Road Traffic Volume Malaysia (RTVM). In 1982, the first printed copy of the organized national traffic census was published by the RTVM. From early 1999, the data were available in Compact Disc (CD). In contrast, the online version started from 2014 to date. The traffic sensor is conducted twice a year during March-April or September-October. The data collection is categorized into three types, type 0, type 1, and type 2. The type 0 data is recorded for 24 hours in 7 days, while type

1 for 16 hours in 7 days, and type 2 for 16 in 1 day. The recording for 16 hours started from 6.00 am to 10.00 pm. The vehicles are divided into six classes, as presented in Table I. We also utilize the air pollution dataset from the AQM stations provided by the Department of Environment (DOE), Malaysia and feature engineered in our previous work [11, 25, 26]

TABLE I. VEHICLE CLASSES

No	Class	Types of Vehicle
1	Class 1	Motor Cars
		Taxi
2	Class 2	Small Vans
		Utilities (Light 2-axles)
3	Class 3	Lorries
		Large Vans (Heavy 2-axles)
4	Class 4	Lorries with 3-axles and above
5	Class 5	Buses
6	Class 6	Motorcycles
		Scooters

The RTVM divides the carriageway into two, single and dual carriageway. Table II describes the carriageway types, their code, and how many lanes each road includes.

TABLE II. TYPE OF CARRIAGEWAY

Code	Single Carriageway	Code	Dual Carriageway
T1-1	Two-Lane	**K1+2**	Three Lane - One way
T1-2	Three-Lane	**K2+2**	Four Lane - One way
T2	Two-Lane - One way	**K2-2**	Four Lane
T2-2	Four-Lane	**k(1+2)-(1+2)**	Six Lane
T3	Three-Lane - One way	—	

Fig. 3 summarizes the total number of traffic census stations in each region in Malaysia. Johor is recorded as the highest region with 75 stations, while Kuala Lumpur with the lowest number of stations with 5.

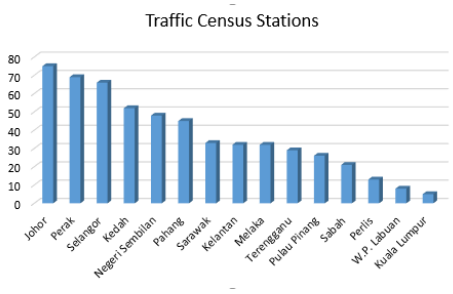

Fig. 3. Traffic Census Stations in Malaysia.

In this study, we proposed a new algorithm for estimating hourly traffic data based on AADT data provided by RTVM. Kuala Lumpur traffic census station was chosen for this study, and it has five stations; these stations are dual carriages with six lanes. Table III presents the station's ID, locations, and a kilometer of each road. The WR101 station was chosen in this study. The traffic data for 2014, 2015, and 2016 were used in this study. Table IV summarizes the hourly (one peak hour) and daily (16 hours) data based on the average estimation given by RTVM.

TABLE III. KUALA LUMPUR CENSUS STATION

Station ID	Location	Kilometer
WR101	Kuala Lumpur - Kuala Selangor (Jalan Kepong)	8.9
WR102	Kuala Lumpur - Ipoh	12.1
WR103	Kuala Lumpur Ipoh	8.1
WR105	Kuala Lumpur - Seremban Expressway	8.1
WR106	Kuala Lumpur - Damansara	5.8

TABLE IV. TRAFFIC VOLUME FOR THREE YEARS

Year	Month	Daily	Hourly	Time
2014	March	133180	9415	9-10
	September	131623	10979	8-9
2015	March	119921	9750	10-11
	September	102273	8031	8-9
2016	April	132029	10083	10-11
	October	113347	7787	18-19

B. Feature Engineering Algorithms

1) Feature Estimation: The feature engineering algorithm is proposed to estimate hourly traffic volume. The algorithm performs two tasks; EstimationOfData and DataDistribution. The EstimationOfData is performed by selecting the station, peak hour, and the normal hour. Six peak hours (peak hours 7.00 to 8.00 am, 8.00 to 9.00 am, 9.00 to 10.00 am, 10.00 to 11.00 am, 17.00 to 18.00 pm, and 18.00 to 19.00 pm) were selected and distributed randomly from the daily average traffic volume for six months, while the remaining amount of the daily traffic volume were distributed randomly to the ten hours (normal hours 11.00 am to 12.00 pm, 12.00 to 13.00 pm, 13.00 to 14.00 pm, 14.00 to 15.00 pm, 15.00 to 16.00 pm, 16.00 to 17.00 pm, 19.00 to 20.00 pm, 20.00 to 21.00 pm, 21.00 to 22.00 pm, and 22.00 to 23.00 pm) for six months as given in the following equations:

$$p = h \tag{1}$$

$$n = d - p \tag{2}$$

The p is the peak hour, h stand for hourly data given by RTVM, n is the normal hour, d daily traffic volume. The percentage of the vehicle type was distributed from the total daily traffic volume using the below equation. The v is the vehicle type, d is the daily traffic volume, and c is the percentage of vehicle type.

$$v = d * p \tag{3}$$

The DataDistribution is the distribution of the estimated traffic data obtained from EstimationOfData. Since the data is based on a six-month average, we create three-years data with hourly rows for sixteen hours, because the RTVM data is based on sixteen hours. We first distribute the amount of peak

TABLE V. PREVIOUS STUDIES THAT USED RTVM DATASET

Author	Objective	Location	Dataset
[27]	Condition of vehicle engine based on driving behavior	Kuala Lumpur	RTVM_2015
[28]	Effect of traffic-related air pollution on traffic policemen	Klang Valley	RTVM_2017
[29]	Association between traffic-related air pollution with respiratory symptoms and DNA damage	Kajang, Hulu Langat, Selangor	RTVM_2016
[30]	Proposed an intelligent national transportation management center	Kuala Lumpur	RTVM_2016
[31]	The autonomous emergency braking system was proposed for the primary accident that is occurring	Jalan Butterworth, Penang	RTVM_2016
[32]	Proposed preventive model for road maintenance	Petaling	RTVM
[33]	Exploring the effect contributing to the vehicle accident	Sabah	RTVM
[34]	Investigating the influence of rubbernecking towards vehicle deceleration rate due to primary accident in the urban area	Jalan Butterworth and Jalan George Town, Penang	RTVM_2016
[31]	Proposed Malaysia driving cycle (MDC) for light-duty test	Terengganu	RTVM_2015
[35]	Investigates driving cycle which contributes to producing air pollution and fuel consumption	Kuala Lumpur	RTVM
[36]	Estimation of particulate matter from non-exhaust and exhaust vehicle	Klang Valley	RTVM_2014
[34]	Development of the driving cycle for route selection	Penang	RTVM_2015
[37]	Investigates single-vehicle accident along with mountainous areas	Sabah	RTVM_2013
[38]	The concentration of air pollution near a primary schools	Pahang	RTVM_2014
[39]	Driving behavior among heavy truck drivers due to pavement damage	Ampang, Kuala Lumpur	RTVM
[40]	Proposed a model for estimating annual daily traffic for single carriageway	Johor	RTVM_2012
[41]	Level of CO and CO_2 due to the increase of motor vehicles in industrial areas	Shah Alam, Seremban, and Kuantan	RTVM
[42]	Estimation of CO produced by passenger cars	Selangor	RTVM_2011
[43]	Vehicle compassion in Malaysia and Indonesia	Malaysia	RTVM_2009
[44]	Estimation of future traffic volume	Skudai, Johor	RTVM_2011
[45]	The trend of public transport in Perak	Perak	RTVM_2005

hour for the six hours randomly and then insert and distribute the normal hours randomly (the remaining ten hours), which is the remaining ten hours. Lastly, we distribute the amount to the type of vehicles based on the percentage given in RTVM data. The algorithm is presented in Fig. 4. The six peak hours were distributed randomly using range with minimum and maximum values (so that the values would not be same). Similarly, normal hours were distributed randomly using range with minimum and maximum values.

2) Feature Generation: In this study, queuing theory is applied to calculate average speed of vehicle and time spent on the road. Queuing theory is a mathematical study of estimation of the waiting time in the queue. Queuing theory is a mathematical study of estimation of the waiting time in the queue. Queuing system considers arrival time, number of server and service time. The arrival time is considered as the arrival of motor vehicles on the particular road segment, the server is the installed camera that recorded the passing vehicles, while the service rate is the time after the vehicle leaves where the camera was installed. The Queuing model structure is presented in Fig. 5, and the notations used in the model.

The arrival rate of the vehicles is distributed using poison distribution. Similarly, the service rate is exponentially distributed. The time spent or waiting time is calculated using the following equation:

$$w = \frac{\lambda^2}{\mu(\mu - \lambda)} \tag{4}$$

Algorithm 1: EstimationOfData&DataDistribution

```
Input: Daily Traffic Volume
Output: Hourly Traffic Volume Estimation
Data: Peak Hour p
foreach Six_month do
    if p = h then
        for _ in range(p):
        value = randit(min, max)
        print(value)
        /* The p is the peak hour, h stand for hourly data given by RTVM.
           min is minimum and max is the maximum                          */
    else if != p then
        n = d - p
        for _ in range(n):
        value = randit(min, max)
        print(value)
        /* The n is the normal hour, d daily traffic volume               */
foreach type_of_vehicle do
    if v = c then
        /* The v is the type of vehicle, d is the daily traffic volume,
           and c is the percentage of vehicle type                        */
        v ++
    else if != v then
        END
END
```

Fig. 4. Feature Estimation Algorithm.

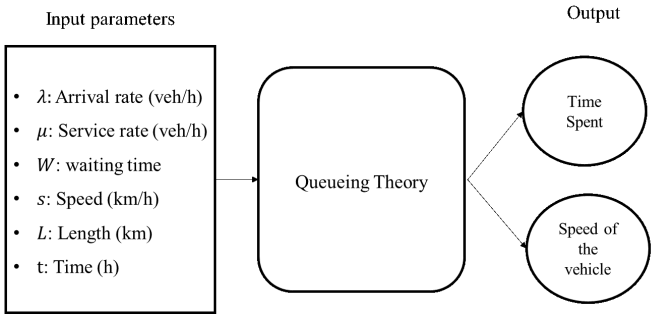

Fig. 5. Queuing Theory & Notations.

The average speed of the motor vehicles is calculated using the below equation:

$$s = \frac{L}{t} \tag{5}$$

IV. Result

A. Feature Engineering Algorithms

Motor vehicle has a significant impact on traffic emissions concentration, and human health; moreover, it causes accidents, traffic congestion, and fuel consumption. The emergence of an automatic traffic management system allows us to record and monitor every vehicle passing on the road. The recorded data are used primarily in transportation studies. In Malaysia, the RTVM provided annual average traffic data. There is a need to estimate the hourly traffic volume. This study proposed a feature engineering algorithm that will efficiently estimate hourly traffic volume and generate features from the existing dataset in Malaysia's traffic census stations. The RTVM gives an average hourly (for peak hour) and daily traffic volume in specific stations. The proposed algorithm was able to estimate the hourly traffic volume for three years based on the yearly average provided by RTVM and distributed the types of vehicles based on the percentage given in the data. Furthermore, the queuing theory was able to generate the vehicle's average speed and time spent on the road. The output of the feature engineering algorithms (estimated and generated features) was shown in Fig. 6.

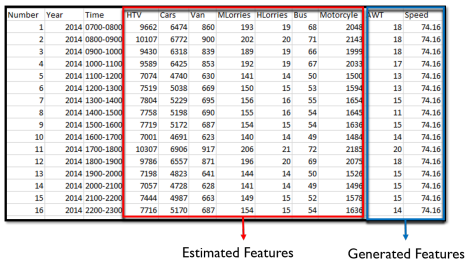

Fig. 6. Feature Engineering Algorithms Result.

Fig. 6 shows the estimated features from the left, which was highlighted with red color. The estimated features were Hourly Traffic Volume (HTV), and types of vehicles (Car and Taxi, Van and Utilities, Medium Lorries, Heavy Lorries, Buses, and Motorcycles) and generated feature from the right with blue color highlighted (Waiting time on the road and average speed of the vehicle).

B. Prediction of Traffic Emission With and Without Traffic Dataset

Due to the lack of studies that generate and estimate features from the RTVM dataset. To justify the claim that feature engineering improves machine learning models' performance, we proposed Random Forest and Decision Three machine learning algorithms to predict traffic emissions concentrations using the estimated and generated features (traffic dataset). Additional dataset of air quality and meteorological variables in [11] study were used. The input and output variables were presented in Table VI.

TABLE VI. Input& Output Features

Input Variables		Output Variables
Traffic Variables	**Meteorological Variables**	**Air Pollutants**
Traffic Volume, Types of vehicle (Car and Taxi, Van and Utilities, Medium Lorries, Heavy Lorries, Buses, and Motorcycles.	Wind Speed, Wind Direction, Temperature, and Relative Humidity.	Carbon Monoxide (CO), Nitrogen Monoxide(NO), Nitrogen Dioxide (NO_2), Nitrogen Oxides (NO_x).

Evaluation metrics such as Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) were used to evaluate the performance of the models. First of all, we predict the level of the CO, NO, NO_2, and NO_x pollutants using meteorological features but without traffic data. Lastly, we included the traffic dataset (estimated and generated features) for prediction. The result shows that our feature engineering algorithms improve the accuracy of the machine learning models (Random Forest and Decision Tree models) by predicting the level of traffic pollutants except for the NO_2, which shows no improvement using the Random Forest model as presented in Table **VII** and **VIII**. We can also visualize the results in Figures 7, 8, and 9 for the Random Forest Model, while Figures 10, 11, and 12 for the Decision Tree Model.

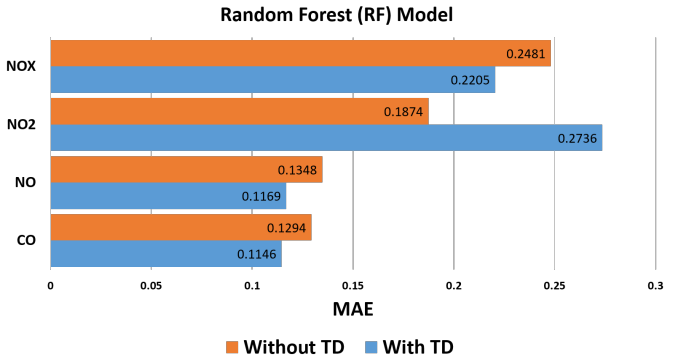

Fig. 7. MAE for the Random Forest Model.

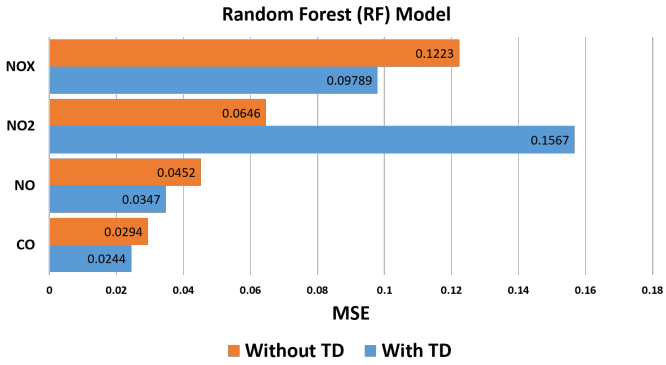

Fig. 8. MSE for the Random Forest Model.

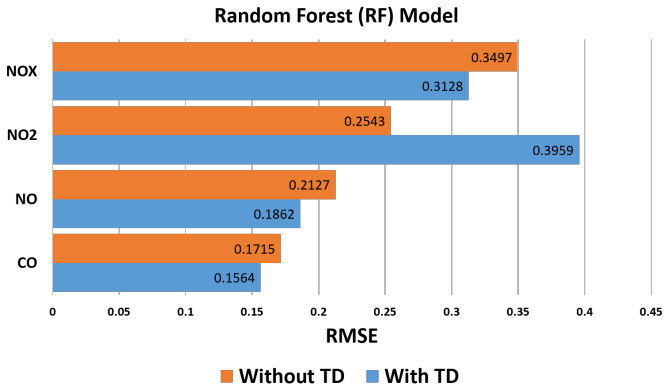

Fig. 9. RMSE for the Random Forest Model.

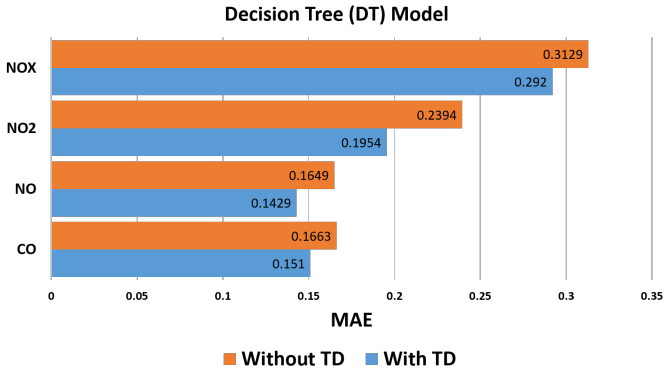

Fig. 10. MAE for the Decision Tree Model.

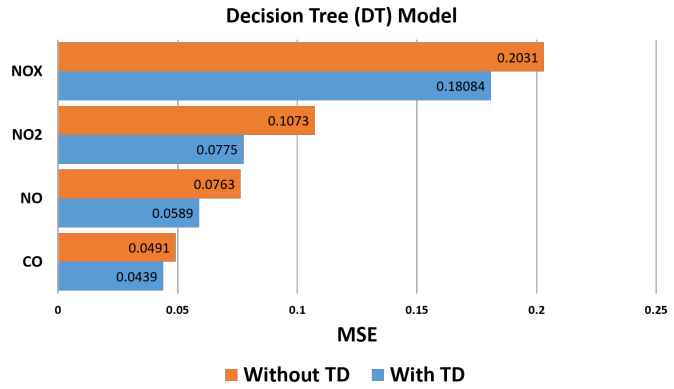

Fig. 11. MSE for the Decision Tree Model.

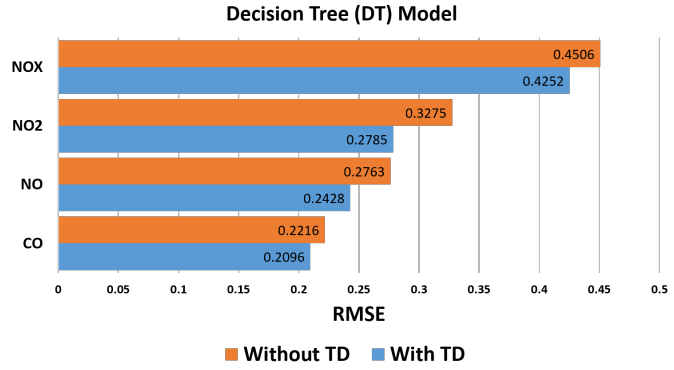

Fig. 12. RMSE for the Decision Tree Model.

TABLE VII. PREDICTION WITH AND WITHOUT TRAFFIC DATA USING RANDOM FOREST MODEL

Random Forest (RF) Model						
Without Traffic Data				With Traffic Data		
Pollutant	MAE	MSE	RMSE	MAE	MSE	RMSE
CO	0.1294	0.0294	0.1715	0.1146	0.0244	0.1564
NO	0.1348	0.0452	0.2127	0.1169	0.0347	0.1862
NO_2	0.1874	0.0646	0.2543	0.2736	0.1567	0.3959
NO_x	0.2481	0.1223	0.3497	0.2205	0.09789	0.3128

TABLE VIII. PREDICTION WITH AND WITHOUT TRAFFIC DATA USING DECISION TREE MODEL

Decision Tree (DT) Model						
Without Traffic Data				With Traffic Data		
Pollutant	MAE	MSE	RMSE	MAE	MSE	RMSE
CO	0.1663	0.0491	0.2216	0.151	0.0439	0.2096
NO	0.1649	0.0763	0.2763	0.1429	0.0589	0.2428
NO_2	0.2394	0.1073	0.3275	0.1954	0.0775	0.2785
NO_x	0.3129	0.2031	0.4506	0.292	0.18084	0.4252

Fig. 7, 10, 8, 11, 9, and 12 show that the performance of RF model for prediction with traffic data were better than the DT with lower MAE, MSE, and RMSE for CO, NO, and NO_x, except NO_2 which show higher MAE, MSE, and RMSE. Similarly, for the prediction without traffic dataset, the RF performed better than DT with lower MAE, MSE and

RMSE for CO, NO, NO_2 and NO_x. We can conclude that the Random Forest Model was a good choice for the prediction of traffic emission concentrations using traffic dataset and without traffic dataset also.

V. Discussion

Motor vehicles become one of the primary concern for the government and agencies. Automobiles create many problems such accident and emissions of pollution to atmosphere. The air pollution produced by motor vehicles had significant impact on human health and the environment as well. Several studies have been conducted for estimation and prediction of traffic emissions. The variability and increase of air pollution depend on traffic characteristics. One of the issue researchers are facing is when the traffic data was provided as an annual average daily traffic (ADDT). The RTVM provides ADDT dataset. The data were incomplete as shown in the Fig. 1. We proposed feature engineering algorithms to estimated and generate missing values in the RTVM dataset. Our feature engineering algorithms were able to generate and estimate the missing features as presented in the Fig. 6. Our feature engineering algorithm is applied in one of the stations in Kuala Lumpur traffic census stations, and this algorithm can be used on the other stations in Malaysia. Additionally, the algorithm can also be used for any annual average daily traffic data if it includes an average hourly dataset. There some limitations in this study, firstly, the estimated hourly traffic volume is proposed due to insufficient hourly traffic volume, which may not provide the exact traffic volume hourly. The RTVM does not provide speed of the vehicle, we calculate vehicle speed as an average basis (the speed of the vehicle is constant). This study could not extract acceleration/deceleration. Some studies suggested that different types of fuel have different emissions, but consideration of fuel type is not provided in this study. Jalan Kepong traffic census station was the selected station in this study, the remaining stations were not studied.

VI. Conclusion

Motor vehicles are the primary source of air pollution in metropolitan globally. Air pollution has a significant effect on human health with diseases such as asthma, cardiovascular, and respiratory. Motor vehicle also causes accidents and create congestion at road segments. Due to these reasons, the government and agencies introduce an automated traffic management system to record the passing vehicles on the road. The recorded data has been used in various studies by researchers. The Road Traffic Volume Malaysia provides incomplete traffic data. We proposed a new feature engineering algorithm to overcome the issue of incomplete traffic data. The proposed feature engineering algorithms could estimate the hourly traffic volume and generate features for three years in Jalan Kepong, Kuala Lumpur, Malaysia. The algorithm was evaluated by predicting four traffic pollutants CO, NO, NO_2, and NO_x using Random Forest and Decision Tree models. The prediction was conducted in two phases, phase one is prediction without traffic dataset (estimated and generated features), and phase two is the prediction with traffic dataset. The result shows that our feature engineering algorithms improve machine learning models' performance except for the prediction of NO_2 using Random Forest, which shows the highest MAE, MSE, and RMSE when traffic data was included for prediction.

Acknowledgment

This research is funded by Taylor's University under the research grant application ID (TUFR/2017/004/04) entitled as "Modeling and Visualization of Air-Pollution and its Impacts on Health". We are also thankful to Department of Environment, Malaysia for providing the AQM station dataset and Ministry of Works, Malaysia for providing the RTVM dataset.

References

[1] R. S. A. Usmani, A. Saeed, A. M. Abdullahi, T. R. Pillai, N. Z. Jhanjhi, and I. A. T. Hashem, "Air pollution and its health impacts in Malaysia: a review," *Air Quality, Atmosphere & Health*, jul 2020. [Online]. Available: https://doi.org/10.1007/s11869-020-00867-xhttp://link.springer.com/10.1007/s11869-020-00867-x

[2] R. S. A. Usmani, I. A. T. Hashem, T. R. Pillai, A. Saeed, and A. M. Abdullahi, "Geographic Information System and Big Spatial Data," *International Journal of Enterprise Information Systems (IJEIS)*, vol. 16, no. 4, 2020.

[3] M. Bilal, R. S. A. Usmani, M. Tayyab, A. A. Mahmoud, R. M. Abdalla, M. Marjani, T. R. Pillai, and I. A. Targio Hashem, "Smart Cities Data: Framework, Applications, and Challenges BT - Handbook of Smart Cities," J. C. Augusto, Ed. Cham: Springer International Publishing, 2020, pp. 1–29. [Online]. Available: https://doi.org/10.1007/978-3-030-15145-4{_}6-1

[4] R. Rossi, R. Ceccato, and M. Gastaldi, "Effect of road traffic on air pollution. experimental evidence from covid-19 lockdown," *Sustainability*, vol. 12, no. 21, p. 8984, 2020.

[5] P. Krecl, Y. A. Cipoli, A. C. Targino, L. B. Castro, L. Gidhagen, F. Malucelli, and A. Wolf, "Cyclists' exposure to air pollution under different traffic management strategies," *Science of the Total Environment*, vol. 723, p. 138043, 2020.

[6] N. Abdull, M. Yoneda, and Y. Shimada, "Traffic characteristics and pollutant emission from road transport in urban area," *Air Quality, Atmosphere & Health*, vol. 13, no. 6, pp. 731–738, 2020.

[7] S. S. Anjum, R. M. Noor, N. Aghamohammadi, I. Ahmedy, L. M. Kiah, N. Hussin, M. H. Anisi, and M. A. Qureshi, "Modeling traffic congestion based on air quality for greener environment: an empirical study," *IEEE Access*, vol. 7, pp. 57 100–57 119, 2019.

[8] O. V. Kurnykina, O. V. Popova, S. V. Zubkova, D. V. Karpukhin, V. P. Pavlov, P. K. Varenik, I. A. Aleshkova, and L. Y. Novitskaya, "Air pollution by road traffic and its measurement methods," *EurAsian Journal of BioSciences*, vol. 12, no. 2, pp. 181–188, 2018.

[9] V. Kumar, J. Prasad, and B. Singh, "Traffic density estimation using progressive neural architecture search," *Journal of Statistics and Management Systems*, vol. 23, no. 2, pp. 481–493, 2020.

[10] A. Sfyridis and P. Agnolucci, "Annual average daily traffic estimation in england and wales: An application of clustering and regression modelling," *Journal of Transport Geography*, vol. 83, p. 102658, 2020.

[11] R. S. A. Usmani, W. N. F. B. W. Azmi, A. M. Abdullahi, I. A. T. Hashem, and T. R. Pillai, "A novel feature engineering algorithm for air quality datasets," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 3, sep 2020.

[12] Z. Yi, X. C. Liu, N. Markovic, and J. Phillips, "Inferencing hourly traffic volume using data-driven machine learning and graph theory," *Computers, Environment and Urban Systems*, vol. 85, p. 101548, 2021.

[13] Z. Dai, H. Song, H. Liang, F. Wu, X. Wang, J. Jia, and Y. Fang, "Traffic parameter estimation and control system based on machine vision," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2020.

[14] S. Zahedian, P. Sekuła, A. Nohekhan, and Z. Vander Laan, "Estimating hourly traffic volumes using artificial neural network with additional inputs from automatic traffic recorders," *Transportation Research Record*, vol. 2674, no. 3, pp. 272–282, 2020.

[15] C. Zhang, S. Shen, H. Huang, and L. Wang, "Estimation of the vehicle speed using cross-correlation algorithms and mems wireless sensors," *Sensors*, vol. 21, no. 5, p. 1721, 2021.

[16] C.-H. Chen, "A cell probe-based method for vehicle speed estimation," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 103, no. 1, pp. 265–267, 2020.

[17] Z. Zhang and X. Yang, "Freeway traffic speed estimation by regression machine-learning techniques using probe vehicle and sensor detector

data," *Journal of transportation engineering, Part A: Systems*, vol. 146, no. 12, p. 04020138, 2020.

[18] L. R. Costa, M. S. Rauen, and A. B. Fronza, "Car speed estimation based on image scale factor," *Forensic science international*, vol. 310, p. 110229, 2020.

[19] Y. Feng, G. Mao, B. Cheng, C. Li, Y. Hui, Z. Xu, and J. Chen, "Mag-monitor: Vehicle speed estimation and vehicle classification through a magnetic sensor," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[20] G. Sil, S. Nama, A. Maji, and A. K. Maurya, "Effect of horizontal curve geometry on vehicle speed distribution: a four-lane divided highway study," *Transportation letters*, vol. 12, no. 10, pp. 713–722, 2020.

[21] C.-H. Yang and H.-M. Tsai, "Vehicle counting and speed estimation with rfid backscatter signal," in *2019 IEEE Vehicular Networking Conference (VNC)*. IEEE, 2019, pp. 1–8.

[22] F. Afifah, S. Nasrin, and A. Mukit, "Vehicle speed estimation using image processing," *Journal of Advanced Research in Applied Mechanics*, vol. 48, no. 1, pp. 9–16, 2018.

[23] H. Dong, M. Wen, and Z. Yang, "Vehicle speed estimation based on 3d convnets and non-local blocks," *Future Internet*, vol. 11, no. 6, p. 123, 2019.

[24] A. Kumar, P. Khorramshahi, W.-A. Lin, P. Dhar, J.-C. Chen, and R. Chellappa, "A semi-automatic 2d solution for vehicle speed estimation from monocular videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 137–144.

[25] R. S. A. Usmani, T. R. Pillai, I. A. T. Hashem, N. Z. Jhanjhi, and A. Saeed, "A Spatial Feature Engineering Algorithm for Creating Air Pollution Health Datasets," 2020. [Online]. Available: https://www.techrxiv.org/articles/preprint/A{_}Spatial{_}Feature{_}Engineering{_}Algorithm{_}for{_}Creating{_}Air{_}Pollution{_}Health{_}Datasets/12376427/2

[26] ——, "A Spatial Feature Engineering Algorithm for Creating Air Pollution Health Datasets," nov 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666307420300115https://www.techrxiv.org/articles/preprint/A{_}Spatial{_}Feature{_}Engineering{_}Algorithm{_}for{_}Creating{_}Air{_}Pollution{_}Health{_}Datasets/12376427/2

[27] M. F. M. Suhaimi, N. A. M. Salleh, M. S. Sarip *et al.*, "Development of kuala lumpur driving cycle for the estimation of fuel consumption and vehicular emission," in *IOP Conference Series: Materials Science and Engineering*, vol. 834, no. 1. IOP Publishing, 2020, p. 012040.

[28] N. F. M. Fandi, J. Jalaludin, M. T. Latif, H. H. Abd Hamid, M. F. Awang *et al.*, "Btex exposure assessment and inhalation health risks to traffic policemen in the klang valley region, malaysia," *Aerosol and Air Quality Research*, vol. 20, no. 9, pp. 1922–1937, 2020.

[29] I. N. Ismail, J. Jalaludin, S. A. Bakar, N. H. Hisamuddin, and N. F. Suhaimi, "Association of traffic-related air pollution (trap) with dna damage and respiratory health symptoms among primary school children in selangor." *Asian Journal of Atmospheric Environment (AJAE)*, vol. 13, no. 2, 2019.

[30] N. S. Musa, N. M. M. Noor, and J. M. Marjan, "The benefits of national intelligent transportation management centre (nitmc) establishment in malaysia," in *IOP Conference Series: Materials Science and Engineering*, vol. 512, no. 1. IOP Publishing, 2019, p. 012013.

[31] M. Rani, A. Mahayadin, A. Shahriman, Z. Razlan, I. Zunaidi, W. Wan, A. Harun, M. Hashim, I. Ibrahim, N. Kamarrudin *et al.*, "Terengganu routes representation for development of malaysia driving cycle: Route selection methodology," in *IOP Conference Series: Materials Science and Engineering*, vol. 429, no. 1. IOP Publishing, 2018, p. 012050.

[32] R. Hamsan, H. Hafiz, A. Azlan, M. Keprawi, A. Malik, A. Adamuddin, A. Abdullah, and A. Shafie, "Pavement condition assessment to forecast maintenance program on jkr state roads in petaling district," in *AIP Conference Proceedings*, vol. 1930, no. 1. AIP Publishing LLC, 2018, p. 020021.

[33] R. Rusli, M. M. Haque, A. P. Afghari, and M. King, "Applying a random parameters negative binomial lindley model to examine multi-vehicle crashes along rural mountainous highways in malaysia," *Accident Analysis & Prevention*, vol. 119, pp. 80–90, 2018.

[34] A. Mahayadin, A. Shahriman, M. Hashim, Z. Razlan, M. Faizi, A. Harun, N. Kamarrudin, I. Ibrahim, M. Saad, M. Rani *et al.*, "Efficient methodology of route selection for driving cycle development," in *Journal of Physics: Conference Series*, vol. 908, no. 1. IOP Publishing, 2017, p. 012082.

[35] R. Tharvin, N. Kamarrudin, A. Shahriman, I. Zunaidi, Z. Razlan, W. Wan, A. Harun, M. Hashim, I. Ibrahim, M. Faizi *et al.*, "Development of driving cycle for passenger car under real world driving conditions in kuala lumpur, malaysia," in *IOP Conference Series: Materials Science and Engineering*, vol. 429, no. 1. IOP Publishing, 2018, p. 012047.

[36] R. E. Elhadi, A. M. Abdullah, A. H. Abdullah, Z. H. Ash'aari, D. Gumel, M. A. Jamalani, L. K. Chng, and F. M. Binyehmed, "A gis-based emission inventory at 1 km-1km spatial resolution for particulate matter (pm10) in klang valley, malaysia," *Science International*, vol. 29, no. 2, pp. 49–49, 2017.

[37] R. Rusli, M. M. Haque, M. King, and W. S. Voon, "Single-vehicle crashes along rural mountainous highways in malaysia: an application of random parameters negative binomial model," *Accident Analysis & Prevention*, vol. 102, pp. 153–164, 2017.

[38] M. Zahaba, H. Abdul Hadi, H. Ariffin, N. Abdull, N. Samsuddin, and M. S. Mohd Aris, "Composition and source determination of heavy metals (hm) in particles in selected primary schools in pahang," *ESTEEM Academic Journal*, vol. 13, pp. 176–184, 2017.

[39] N. A. Khalifa, A. Alsnose, A. Zulkiple, and R. Z. Abidin, "Non-pragmatic data collection for road pavement damage on access road to residential estate and the statistical analysis choice," *International Journal of Traffic and Transportation Engineering*, vol. 5, pp. 83–90, 2016.

[40] N. S. M. Nor, O. C. Puan, N. Mashros, and M. K. B. Ibrahim, "Estimating average daily traffic using alternative method for single carriageway road in southern region malaysia," *ARPN Journal of Engineering and Applied Sciences*, 2006.

[41] A. AZHARI, A. F. MOHAMED, and M. T. LATIF, "Carbon emission from vehicular source in selected industrial areas in malaysia," *International Journal of the Malay World and Civilisation*, vol. 4, no. 1, pp. 89–93, 2016.

[42] R. E. Elhadi, D. Y. Gumel, and M. Fallah, "Co2 emission inventoryof onroad vehicles in selangor state inpeninsular malaysia," *International Journal of Advanced Scientific and Technical Research*, 2015.

[43] L. S. Putranto, J. Prasetijo, and N. L. P. S. E. Setyarini, "Vehicle composition in indonesia and malaysia," *The 10th Eastern Asia Society for Transportation Studies*, 2013.

[44] A. Minhans, N. H. Zaki, and R. Belwal, "Traffic impact assessment: A case of proposed hypermarket in skudai town of malaysia," *Jurnal Teknologi*, vol. 65, no. 3, 2013.

[45] W. Suwardo, M. Napiah, and I. Kamaruddin, "Review on motorization and use of public transport in perak malaysia: realities and challenges," *2nd INTERNATIONAL CONFERENCE ON BUILT ENVIRONMENT IN DEVELOPING COUNTRIES*, 2008.