Next Energy 1 (2023) 100060

Contents lists available at ScienceDirect

Next Energy

journal homepage: www.sciencedirect.com/journal/next-energy

Research article

Real-driving CO_2, NO_x and fuel consumption estimation using machine learning approaches

G M Hasan Shahariar [a,b,*], Timothy A. Bodisco [a,b], Nicholas Surawski [c], Md Mostafizur Rahman Komol [d,e], Mojibul Sajjad [a,b], Thuy Chu-Van [a,b], Zoran Ristovski [a,b], Richard J. Brown [a,b]

[a] Biofuel Engine Research Facility, QUT, Brisbane, QLD 4000, Australia
[b] International Laboratory for Air Quality and Health, QUT, Brisbane, QLD 4000, Australia
[c] Centre for Green Technology, University of Technology Sydney, 81 Broadway, Ultimo, NSW 2007, Australia
[d] Centre for Robotics, QUT, Brisbane, QLD 4000, Australia
[e] CSIRO, Data61, Robotics and Autonomous System, Pullenvale, QLD 4069, Australia

ARTICLE INFO

Keywords:
Machine learning
Driving behaviour
RDE
Emission
Fuel consumption
PEMS

ABSTRACT

Real driving emissions (RDE) testing are gaining attention for monitoring and regulatory purposes because of providing more realistic emission and fuel consumption measurements compared to laboratory tests. This study aims to develop machine learning (ML) based emission and fuel consumption estimation models using real-driving measurement data. A light-duty diesel vehicle equipped with a portable emissions measurement system (PEMS) was driven in an urban test route by 30 participant drivers of disparate backgrounds to obtain a wide variety of data in terms of driving behaviour and traffic conditions. The Pearson correlation coefficient was used to select the input variables among 36 driving behaviours and 6 engine parameters. The CO_2, NO_x and fuel consumption prediction models were developed using linear regression (LR), support vector machine (SVM) and Gaussian process regression (GPR). The results showed that all three models could predict CO_2 with an absolute relative error (ARE) of less than 9%. The GPR model showed the best performance in CO_2 prediction with an R^2 of 0.74 and ARE of 3.30%. LR model showed the best prediction accuracy for NO_x with an R^2 of 0.80 and ARE of 8.91%. All three models worked well for fuel consumption prediction, however, GPR showed the best accuracy with an R^2 of 0.81 and ARE of 3.52%. This method lays a foundation for developing route/region specific emission and fuel consumption models that will help to monitor and reduce the environmental impact and the amount of burned fuel. Moreover, developing models from different driver classes will provide valuable insights into emission-optimal driving behaviour which could be used to train new drivers.

1. Introduction

Emissions from on-road vehicles are a major source of urban air pollution, where half of the world's population lives [11]. Increasing vehicle population causes traffic congestion and degraded air quality due to an increase in emissions. For example, in Europe, approximately 50% of NO_x and particulate matter are produced by on-road vehicles [4]. Multiple research indicated that on-road emissions increase risks to the morbidity and mortality of road users and nearby communities [21,32,41].

To control air pollution from on-road vehicles, strengthening emission legislation has been gradually adopted by regulatory authorities [20]. These regulations influenced the regulatory authorities to adopt more strict emission certification procedures [12,17]. RDE was introduced by the European Commission as a certification procedure [46] which encompasses several driving factors including driving behaviour, traffic condition, road grade and road environment. Several recent studies have reported significantly higher emissions during on-road measurements than those reported during chassis dynamometer testing [51,55].

Several factors have been identified to affect fuel consumption and emissions. These could be roughly categorised as weather-related (wind speed and direction, humidity and temperature), road topography (road grade, elevation, surface roughness), road environment (traffic

* Corresponding author at: Biofuel Engine Research Facility, QUT, Brisbane, QLD 4000, Australia.
E-mail address: hasanshahariar.kuet@gmail.com (G.M.H. Shahariar).

https://doi.org/10.1016/j.nxener.2023.100060
Received 12 April 2023; Received in revised form 21 August 2023; Accepted 12 September 2023
Available online 19 September 2023

condition, road features, travel distance), driving dynamics (speed, acceleration, deceleration, power demand) and driving behaviour (timid, aggressive) [24,58]. Donateo and Giovinazzi, [15] conducted an RDE study in southern Italy during summer using the exact vehicle and same driver to assess the variability of emissions in terms of weather conditions together with traffic conditions to optimise the route for RDE cycles. This study has found that on-road emissions are strongly affected by ambient conditions together with engine conditions changes due to traffic. It is evident from numerous studies that among the above-mentioned factors, driving behaviour plays an important role in real-driving emissions and fuel consumption [22,48,53]. Huang et al., [27] reviewed different eco-driving studies and summarised that up to 15% fuel savings could be achieved. Pelkmans et al., [42] conducted measurements of a bus in urban traffic and observed that a driving cycle consists of 35% acceleration which is responsible for 70% of fuel consumption and 60–80% of CO, HC and NO_x emissions. May et al., [35] reported a trip with comparatively high acceleration and deceleration significantly increases emissions and fuel consumption. Therefore, modelling emissions and fuel consumption concerning driving behaviour and other associated parameters is necessary for optimisation purposes.

On-road emissions and fuel consumption are to be modelled using the data obtained from real-driving measurements. Among different modelling approaches, machine learning (ML) is becoming popular in emission and fuel consumption modelling for both diesel and gasoline engines [2,13,19,29,31,38]. Hashemi and Clark, [25] developed ML models for specific laboratory-based driving cycles that can predict CO_2 and NO_x with comparatively better accuracy compared to the existing models. Moradi et al., [38] performed ML modelling for highly transient laboratory cycles for gasoline engines. However, modelling real-driving emissions and fuel consumption poses significant challenges due to the non-reproducible nature of driving behaviour and traffic conditions. Yao et al., [56] implemented a backpropagation (BP) neural network, random forest and support vector regression (SVR)-based ML models using driving behaviour and vehicle dynamics data obtained from an onboard diagnostics system (OBD) and mobile phone terminals. These models can predict on-road fuel consumption with an absolute relative error of less than 10%. Moradi and Miranda-Moreno, [37] developed SVM and ANN-based ML models to predict the fuel consumption of a fleet of 27 vehicles using speed, acceleration, road grade and engine speed. Le Cornec et al., [31] developed a look-up table (LT), non-linear regression (NLR) and neural network multi-layer perceptron (MLP) model to predict the instantaneous NO_x using vehicle speed and acceleration.

Several driving dynamics parameters can be derived from a performed drive cycle including total distance, trip time, driving time, cruising time, acceleration time, deceleration time, braking time, idle time, driving percentage, cruising percentage, acceleration percentage, deacceleration percentage, braking percentage, idling percentage, average speed, maximum speed, average acceleration, average positive acceleration, average negative acceleration, number of acceleration, acceleration/km, number of stops, stops/km, average stop duration, relative positive acceleration (RPA), positive kinetic energy (PKE), velocity × positive acceleration (VA), relative positive speed (RPS), relative cubic speed (RCS), root mean square of acceleration (RMSA), urban distance percentage, rural distance percentage, motorway distance percentage, urban trip percentage, rural trip percentage, motorway trip percentage [6]. Each of these parameters has several degrees of correlation with emissions and fuel consumption. Moreover, some of the parameters may have interrelations between them which may have an impact on prediction accuracy. These parameters could be used to analyse and characterise a specific driving cycle and need to be investigated for the development of on-road emissions and fuel consumption modelling. However, as shown in the literature review, very few of these parameters have been studied and considered for the modelling approach [13,19,25,26,31,37,38,45,56].

The selection of input features is vital in ML modelling because it influences prediction results. Fang et al., [19] conducted ML modelling to predict diesel engine NO_x emission. This study reported that excess input parameters can lead to an increase in errors in prediction results, especially low NO_x generating points. To reduce the number of inputs and select the most appropriate input features, the significance level (p-value) and Pearson correlation coefficient were used. However, the p-value was found to be partially misleading in filter-based feature selections, and the Pearson coefficient worked well for the reduction of input variables.

An extensive search of the literature indicates that the majority of RDE modelling studies focused on tailpipe emissions which are treated by after-treatment systems and may not provide the actual correlation with driving dynamics [2,13,19,25,31,37,38,45,50]. To overcome this issue the current study aimed to capture emission data before the after-treatment system, which will directly provide actual engine emission data in correlation with driving dynamics. The sensors were located closer to the engine, rather than at the exit of the tailpipe, to ensure the exhaust gas had as little influence as possible from the after-treatment system.

Existing RDE studies performed using either a single or a small number of drivers, where the drivers have been provided prior training to simulate different driving styles, thereby no longer representing real-world driving [10,20,22,23,30,33,36,44,51,53]. Also, very few studies have modelled RDE emissions and fuel consumption focusing on the high transient operation (urban driving) which contributes a major portion of emissions.

The current study addressed the above-mentioned research gaps and developed ML-based emission and fuel consumption estimation focusing on the following:

- Captured 60 trip data using 30 drivers from various backgrounds in terms of nationality, occupation, gender, driving experience and age. Participating drivers were given no prior training which enabled the capture of data from actual real-world drivers.
- All the emission parameters have been captured before the after-treatment system to represent the raw emissions.
- An urban test route was designed which consists of typical route features such as traffic lights, roundabouts, speed bumps motorway entrances and exits, city centre, school zone etc.

Three different ML models based on LR, SVM and GPR were developed, using driving behaviour and engine-related parameters to predict on-road CO_2, NO_x and fuel consumption, respectively. This study will contribute to more realistic emission and fuel consumption predictions in real-driving conditions which will contribute to developing a real-time monitoring protocol and emission control management planning.

2. Data collection (on-road measurement campaign)

A Hyundai iLoad van was used for the measurements. Couriers and tradespeople prefer this vehicle as it offers good fuel economy and updated features at a competitive price. A 12.5 km route was designated in the Brisbane metropolitan area as most courier and trade vans are driven in urban areas. The test route covers most of the features that a typical urban driver faces every day. A programmable GPS was used to guide the drivers. The test route map is shown in Fig. 1.

Thirty participant drivers drove the vehicle on the test route and all the driving behaviour and engine-related parameters, emissions and fuel consumption were recorded during each trip. The test was conducted in August 2019 and during day time (9 am to 5 pm). The tests exclude the use of auxiliaries such as air conditioning to maintain consistency among tests. The drivers were selected from a wide variety of nationalities, ages, gender, occupation and driving experiences, which enables the achievement of more realistic data from real-world drivers. All 30

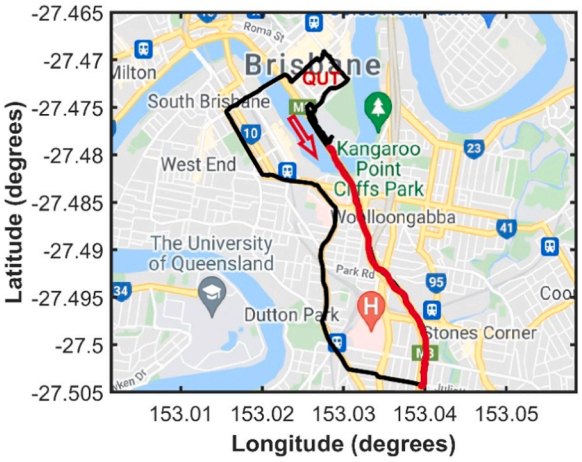

Fig. 1. Designated urban test route.

drivers performed two consecutive trips; hence 60 trip data were collected. A PEMS was used to collect the driving parameters and emissions. The PEMS GPS recorded the instantaneous speed and position of the vehicle. Engine parameters were recorded through the OBDII port, and emission sensors recorded CO_2 and NO_x emission data. An ECM (Engine Control and Monitoring) PEMS was used for the study. Four sampling ports were created 70 cm downstream from the turbocharger along the centreline of the exhaust pipe to install the NO_x, CO_2, pressure, and temperature sensors. The ECM ceramic NO_x sensor was coupled to the pressure compensator kit through the PEMS. The manufacturer stated accuracy is given as $\pm$ 5 ppm (when the NO_x concentration is between 0 and 200 ppm), $\pm$ 20 ppm (when the NO_x concentration is between 200 and 1000 ppm) and $\pm$ 2% (when the NO_x concentration is between 1000 and 5000 ppm). The response time is $<$ 1 s. In-house, MATLAB programs were used for the pre-processing and time alignment of the recorded data. The test vehicle specification is presented in Table 1. Table 2 shows the participant drivers' information.

3. Methodology

3.1. Input parameters

In the first stage, average values of all the variables from the measurements were used as features in the models to examine the trend and behaviour of the data in CO_2, NO_x and fuel consumption estimation. The input features have been classified into two different categories: engine-related parameters and driving behaviour. Engine parameters were recorded using the vehicle OBDII sensors and PEMS sensors. The vehicle speed and time were recorded using the PEMS GPS and all the driving dynamics parameters were calculated using the equations given in [6]. Details of the data acquisition and calculation process of associated engine parameters and driving dynamics parameters are presented in our previous study [49]. The variables of each group are presented in Table 3.

Table 1
Test vehicle specification.

Vehicle	Hyundai iLoad 2017
Fuel	Commercial Diesel
Engine Capacity	2.5 L
Cylinders	4
Maximum torque	441 Nm
Maximum power	125 kW
Odometer	14597 km

Table 2
Participant drivers' information.

Number of Participants	30 (70% male, 30% female)			
Participants' Nationality	70% domestic, 30% international			
Driving Experience (years)	0–5	6–10	11–20	20+
	7%	23%	27%	43%
Drivers' Age (years)	21–30	31–40	41–50	60+
	24%	33%	17%	26%

Table 3
Input variables used for the current study.

Group	Variables
Engine-related parameters	Engine speed, load, throttle, air-fuel ratio (AFR), exhaust flow rate, O_2
Driving behaviour	Total distance, trip time, driving time, cruising time, acceleration time, deceleration time, braking time, idle time, driving percentage, cruising percentage, acceleration percentage, deacceleration percentage, braking percentage, idling percentage, average speed, maximum speed, average acceleration, average positive acceleration, average negative acceleration, number of acceleration, acceleration/km, number of stops, stops/km, average stop duration, RPA, PKE, VA, RPS, RCS, RMSA, urban distance percentage, rural distance percentage, motorway distance percentage, urban trip percentage, rural trip percentage, motorway trip percentage

3.2. Data pre-processing

Pearson correlation is a useful correlation coefficient in statistics that is used to express the relationship between two sets of variables measured in the same interval. It is an effective filter-based feature selection method commonly used in regression analysis and was calculated for all the input parameters for CO_2, NO_x and fuel consumption respectively. By analysing the Pearson correlation, strongly correlated engine parameters and driving behaviour parameters were selected for each target variable. The selected parameters and their Pearson correlation coefficients are presented in Table 4. After that, the input data set was randomly separated into two different groups named test data (used to test the developed models) and training data (for the model training). Test data was checked for three different combinations (20%, 25% and 30%), and it was found that 25% of test data provides the best performance. The value of 25% test data and 75% training data was also suggested by other studies [9]. The test data set should cover the entire

Table 4
Pearson correlation analysis of CO_2, NO_x and fuel consumption with the engine parameters and driving dynamics.

Variable name	Pearson correlation coefficient		
	CO_2 (g/km)	NO_x (g/km)	Fuel consumption (L/hr)
AFR	-0.576	0.370	$<$ 0.1
Exhaust flow rate (kg/h)	-0.326	0.677	$<$ 0.1
Engine load	$<$ 0.1	-0.174	0.539
Throttle (%)	$<$ 0.1	$<$ 0.1	0.814
O_2	$<$ 0.1	$<$ 0.1	-0.342
Total distance (m)	$<$ 0.1	0.328	$<$ 0.1
Trip duration (s)	0.491	0.273	-0.186
Acceleration time (s)	0.388	0.307	$<$ 0.1
Deceleration time (s)	0.376	0.519	$<$ 0.1
Idle time (s)	0.567	$<$ 0.1	-0.255
Average speed (km/h)	-0.609	$<$ 0.1	0.231
Average acceleration (m/s^2)	$<$ 0.1	-0.512	$<$ 0.1
RPA (m/s^2)	0.419	0.668	$<$ 0.1
VA (m^2/s^3)	$<$ 0.1	0.427	$<$ 0.1
PKE (m/s^2)	0.415	0.655	$<$ 0.1
RCS (m^2/s^2)	-0.465	$<$ 0.1	$<$ 0.1

range of the training data set, to ensure holistic testing of the models. Fig. 2(a), (b) and (c) show the parameters most strongly correlated with CO_2, NO_x and fuel consumption, respectively. It can be observed that the test data set (red symbols) represents suitable coverage within the training data set. In order to avoid overfitting, 5-fold cross-validation was used and repeated three times, which is a widely-used method to avoid overfitting in regression modelling [9]. It is evident from Table 4 that driving dynamics parameters such as average speed, idle, RPA and PKE have a strong correlation with emissions. This study captured emissions data before the after-treatment system to avoid its influence. Ultimately, this approach will provide a more accurate estimation of emissions in relation to driving behaviour.

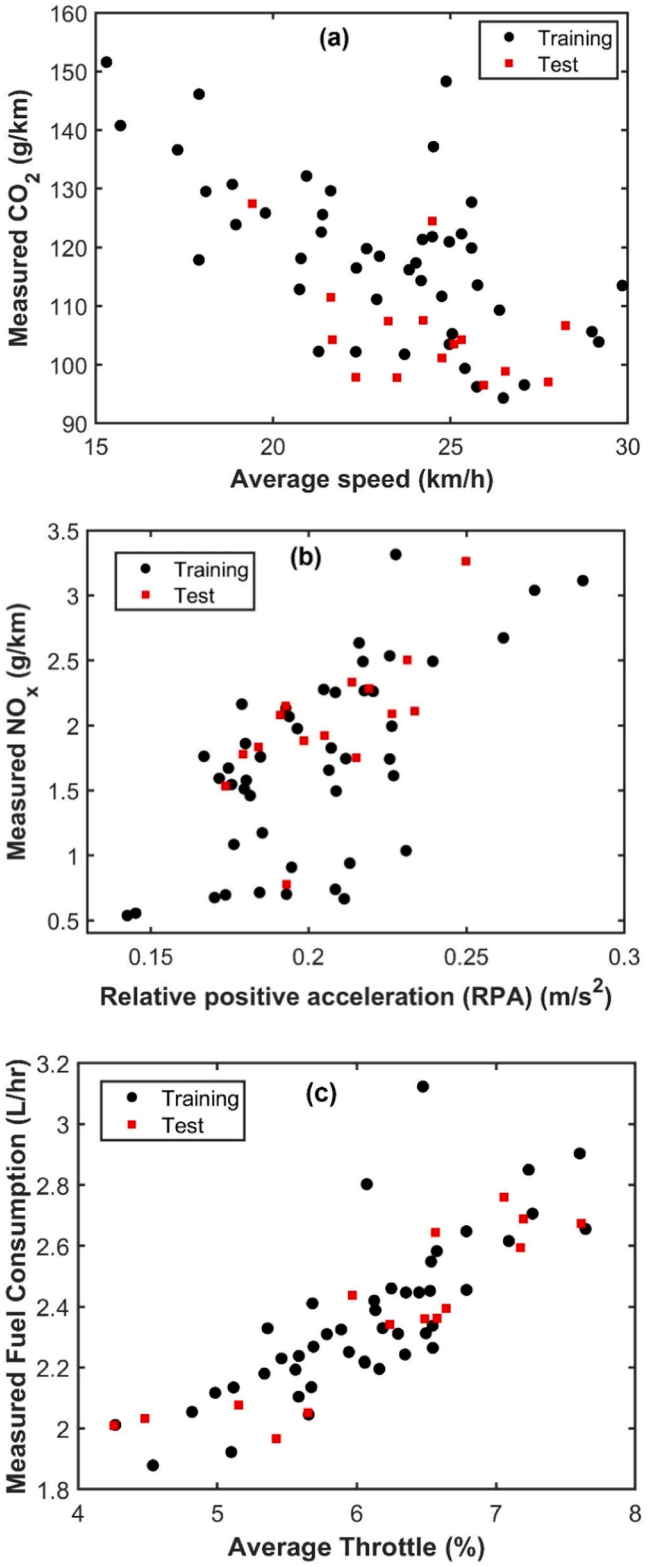

Fig. 2. Illustration of test and training data set for all target parameters: (a) Average speed vs CO_2; (b) RPA vs NO_x; and (c) Throttle (%) vs fuel consumption.

3.3. Multicollinearity diagnostics

A potential problem in multiple regression models that arises from influential observations and high correlations among predictor variables is known as multicollinearity which was first addressed by Belsley et al., [8]. Multicollinearity can influence the slope parameter estimation to have inconsistent magnitude or signs from the expectations or with the bivariate correlation, between a predictor variable and an output variable [52]. Consequently, standard error estimation may increase. Moreover, it can lead to larger confidence intervals which may impact the judgment of the importance of a predictor variable [52]. Therefore, the presence of multicollinearity in the predictor variables of a regression model could provide biased results and incorrect conclusions about the relationship between the output variable and predictor variables. In the presence of multicollinearity, the suggested process by researchers to move forward is to use data reduction techniques i.e., respecifying the regression model with the variables contributing to multicollinearity removed [1,47,52].

Variance inflation factors (VIF) and tolerance are the appropriate parameters for collinearity diagnostics [47,52] and can be calculated as follows:

$$T = 1 - R^2 \tag{1}$$

Here, T is tolerance and R^2 is the coefficient of determination.

$$VIF = \frac{1}{1 - R^2} \tag{2}$$

Here, VIF stands for variance inflation factors.

A tolerance value (T) < 0.1 is considered critical and multicollinearity is present in the predictor data set. In this case, more than 90% of the variance can be explained by the other predictors. In addition, the VIF value increases with increasing multicollinearity and VIF > 10 is considered critical [47,52].

Table 5 presents multicollinearity diagnostics results for the predictor variables associated with CO_2, NO_x and fuel consumption prediction models. Among the CO_2 predictor variables, trip duration and positive kinetic energy (PKE) show significantly higher VIF (45.48 and 428.47, respectively) and significantly lower tolerance (0.022 and 0.002 respectively) than the critical values, hence, they were removed from the prediction model. On the other hand, deceleration time and average speed show very close VIF (11.16 and 12.29, respectively) and tolerance (0.09 and 0.081 respectively) to the critical values and were kept in the prediction models. Among the NO_x predictor variables, acceleration time and positive kinetic energy (PKE) show higher VIF (19.06 and 437.61, respectively) and lower tolerance (0.052 and 0.002, respectively) than the critical values, hence they were removed from the models. All the fuel consumption predictor variables satisfied the critical values of collinearity diagnostics.

3.4. Machine learning approaches

30 drivers participated in the on-road measurement campaign and performed a total of 60 trips on the urban test route. Engine parameters, driving dynamics and emissions were recorded using PEMS. The dataset obtained from real-driving measurements was used for the prediction and engine parameters and driving dynamics were considered as input variables in the machine-learning models. For our analysis of CO_2, NO_x and fuel consumption prediction, five machine learning algorithms were

Table 5
Collinearity statistics of the predictor variables for CO_2, NO_x and fuel consumption prediction models.

Variable name	Collinearity Statistics					
	CO_2 (g/km)		NO_x (g/km)		Fuel consumption (L/hr)	
	Tolerance	VIF	Tolerance	VIF	Tolerance	VIF
AFR	0.480	2.09	0.492	2.03	< 0.05	> 20
Exhaust flow rate (kg/h)	0.184	5.43	0.185	5.41	< 0.05	> 20
Engine load	< 0.05	> 20	0.843	1.18	0.776	1.28
Throttle (%)	< 0.05	> 20	< 0.05	> 20	0.943	1.06
O_2	< 0.05	> 20	< 0.05	> 20	0.835	1.19
Total distance (m)	< 0.05	> 20	0.274	3.64	< 0.05	> 20
Trip duration (s)	0.022	45.48	0.158	6.34	0.205	4.87
Acceleration time (s)	0.111	8.98	0.052	19.06	< 0.05	> 20
Deceleration time (s)	0.090	11.16	0.158	6.32	< 0.05	> 20
Idle time (s)	0.173	5.77	< 0.05	> 20	0.193	5.17
Average speed (km/h)	0.081	12.29	< 0.05	> 20		
Average acceleration (m/s^2)	< 0.05	> 20	0.667	1.50	0.238	4.21
RPA (m/s^2)	0.355	2.816	0.137	7.28	< 0.05	> 20
VA (m^2/s^3)	< 0.05	> 20	0.235	4.25	< 0.05	> 20
PKE (m/s^2)	< 0.05	> 20	< 0.05	> 20	< 0.05	> 20
RCS (m^2/s^2)	0.220	4.55	< 0.05	> 20	< 0.05	> 20

used to train regression models: linear regression, support vector machine, Gaussian processes regression, ensemble regression and random forests. Among these five, two were eliminated because of poor predictive power, leaving three candidate models which were used for the predictions. Among these two, the main drawback of the ensemble method is that it could be prone to overfitting or underfitting if the aggerate method is too complex. Due to a large number of real-world factors associated with on-road emissions, this model could underestimate the uncertainties which are evident from the performance metrics. An important limitation of the random forest model is that it requires a substantially large number of input data which ultimately produces a large number of decision trees for prediction. Therefore, it requires a huge computational time. If the data set is achieved from real-world experiments, it is not always possible to achieve a big data set to create adequate decision trees for this predictive model. In such case risk of underfitting increases significantly, which is evident from the current study.

3.4.1. Linear regression

Linear regression (LR) is a supervised machine learning algorithm that expresses the relationship between a dependent variable and single or multiple independent variables [40]. This process fits a straight line or hyperplane to the input variables, to predict the output variable. This study is using multiple input variables (engine-related parameters and driving behaviour) to predict an output (CO_2, NO_x and fuel consumption) which is known as multivariate linear regression. Multivariate linear regression can be expressed as follows:

$$y = h_a(x) = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n \tag{3}$$

Where, y is the desired output variable (CO_2, NO_x, fuel consumption); x1, x2,….,xn are the input variables (engine parameters and driving dynamics) and a1, a2,….,an are the required coefficients needed to be estimated. This algorithm aims to calculate the coefficient values to ensure the best fit between the predictor variables and the target variable. This can be performed by minimising the objective function. The above hypothesis can be expressed in its vectorised form, along with the objective function as:

$$h_a(x_j) = a^T x_j = \sum_i a_i x_{j,i} \tag{4}$$

$$a = \min \sum_j Cost \quad Function(y_j, h_a) \tag{5}$$

3.4.2. Support vector machine

Support vector machine (SVM) is a supervised machine learning algorithm characterised by the ability to govern the decision function by the use of kernel functions that identify one or multiple separating hyperplanes [16]. This algorithm transforms nonlinear problems into multi-dimensional linear problems with a geometrical explanation, and shows strong resistance to the over-fitting problem and high generalisation performance. With this dataset, the structure of the developed SVM model for CO_2, NO_x and fuel consumption prediction is presented in Fig. 3. Here, $K(x_1,x)$, $K(x_2,x)$…. $K(x_n,x)$ are the kernel functions that transform the nonlinear input variables into higher dimensional liner variables within a decision boundary. Recently, this novel supervised machine learning method has been demonstrated to be a promising prediction tool in the field of intelligent transportation management systems and is widely used in real-world applications. Multiple studies have used this method for real-driving fuel consumption and emission estimation of on-road vehicles [43,56,58].

3.4.3. Gaussian processes regression

Gaussian processes regression (GPR) is a nonparametric, Bayesian approach suitable for machine learning regression applications with higher dimensionality and comparatively small datasets [54]. As a nonparametric approach, GPR estimates the probability distribution of all admissible functions that fit the data, rather than estimating the probability distribution of a specific function. GPR can be expressed by its mean function and covariance function as follows:

$$f(x) \sim GPR((m(x), k(x, \acute{x}) \tag{6}$$

where, f(x) is a function of variable x, m(x) is the main function and k(x, $\acute{x}$) is the covariance function. [54] presented a comprehensive analysis

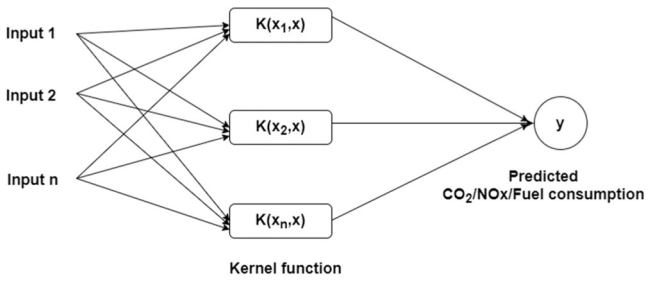

Fig. 3. The structure of on-road CO_2, NO_x and fuel consumption prediction model based on SVM.

of the GPR model in their study. Several studies used the GPR model for ship and aircraft fuel consumption estimation [7,26,57].

3.5. Model performance evaluation

To validate the developed machine learning models and evaluate their accuracy and efficiency, three indices, the root-mean-square error (RMSE), coefficient of determination (R^2) and absolute relative error (ARE) were compared. These are the most important performance evaluation metrics [39] and are calculated as follows:

$$RMSE = \frac{\sqrt{\sum (f_i - y_i)^2}}{n} \tag{7}$$

Here, f_i is the predicted variable and y_i is the measured variable and n is the number of samples.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (f_i - y_i)^2}{\sum_{i=1}^{n} (f_i - \bar{y})^2} \tag{8}$$

Here, f_i are the predicted variables and $\bar{y}$ is the average of the measured variables.

$$ARE = \left| \frac{X_{measured} - X_{predicted}}{X_{measured}} \right| \times 100 \tag{9}$$

Here, $X_{measured}$ are the measured parameters during the on-road test and $X_{predicted}$ are the corresponding estimated parameters.

4. Results and discussion

By applying the developed machine learning regression models (LR, SVM and GPR), CO_2, NO_x and fuel consumption, respectively were predicted using real-driving measurement data. RMSE, R^2 and ARE of all models for CO_2, NO_x and fuel consumption are presented in Table 6. The model evaluation results for CO_2 show that all three models had good predictive power. Among these, GPR and SVM showed greater predictive power. The RMSE is 7.21 g/km and 7.39 g/km and R^2 is 0.74 and 0.72, respectively which represents a higher fitting degree and indicates that these models can accurately predict CO_2 emissions with the data collected from real-driving measurements. The ARE of the predicted CO_2 is significantly lower for GPR compared to SVM and LR. By comparing the key model performance indicators among the three models, it is evident that the GPR has higher accuracy and lower error than the SVM and LR models. Therefore, the CO_2 prediction model based on GPR is effective and efficient for CO_2 emission prediction, based on data obtained from real-driving measurements.

Model performance results show the LR model worked well for NO_x emission prediction with an R^2 value of 0.80, which is significantly higher than that of the SVM and GPR models (0.54 and 0.68, respectively). RMSE is 0.326 g/km and ARE is 8.91, which are also lower than the other two models. In general, the ARE is comparatively higher for NO_x prediction than CO_2 and fuel consumption. This is confirmed by several studies which show that real-driving NO_x emission is significantly unpredictable and depends on several factors [34,44,51]. Overall, the LR model was demonstrated as the most efficient model that can predict NO_x emission from real-driving measurements.

In the case of fuel consumption prediction, the three models all

Table 6
Model evaluation results.

CO_2			
Prediction method	**RMSE**	**R^2**	**ARE**
Linear regression (LR)	8.71	0.62	8.04
Support vector machine (SVM)	7.39	0.72	6.81
Gaussian process regression (GPR)	**7.21**	**0.74**	**3.30**
NO_x			
Prediction method	**RMSE**	**R^2**	**ARE**
Linear regression (LR)	**0.326**	**0.80**	**8.91**
Support vector machine (SVM)	0.496	0.54	11.86
Gaussian process regression (GPR)	0.415	0.68	12.36
Fuel consumption			
Prediction method	**RMSE**	**R^2**	**ARE**
Linear regression (LR)	0.133	0.74	3.72
Support vector machine (SVM)	**0.111**	**0.82**	**3.41**
Gaussian process regression (GPR)	0.113	0.81	3.52

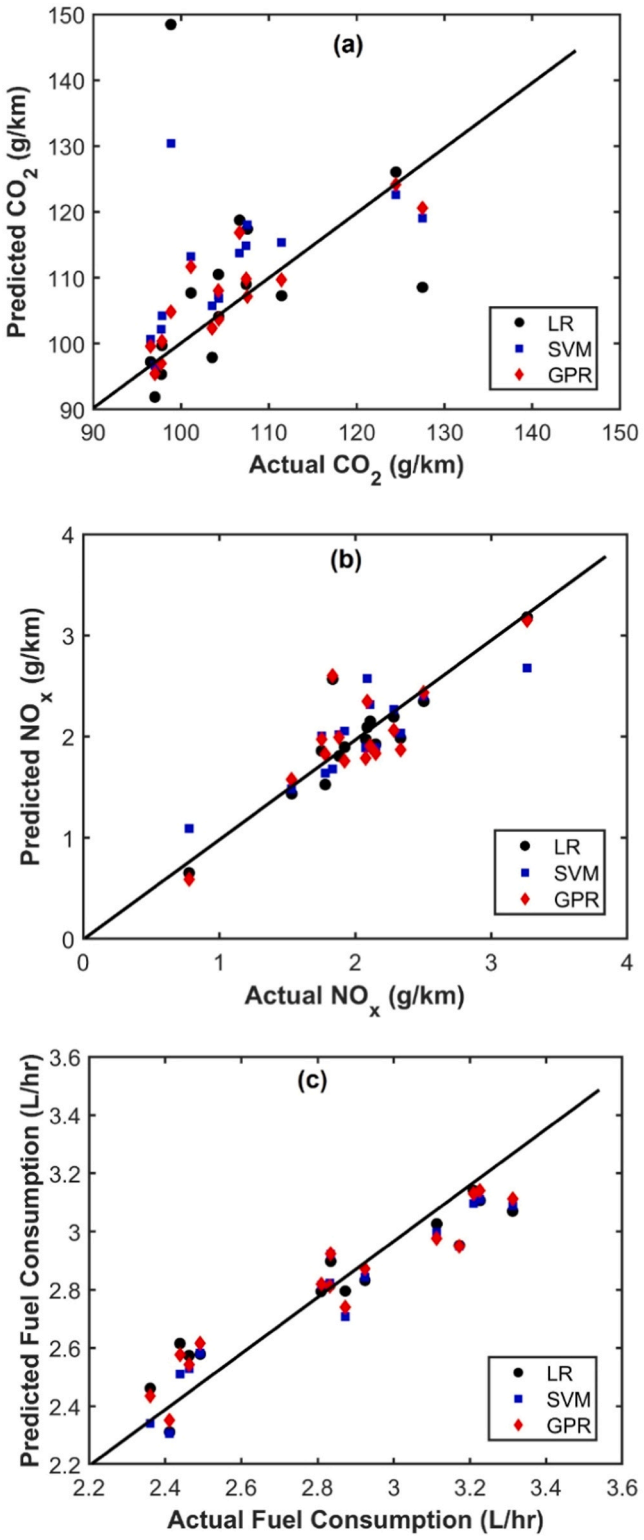

Fig. 4. Prediction results of the three models (a) CO_2 emission, (b) NO_x emission, (c) fuel consumption.

showed high prediction accuracy. In a real-driving scenario vehicle, fuel consumption strongly depends on engine load and throttle, evident from the Pearson correlation coefficient. These two parameters have a strong influence on fuel consumption prediction. Among these three models, SVM showed the highest R^2 value (0.82), and the lowest RMSE (0.111 L/hr) and ARE (3.41). Therefore, an SVM-based fuel consumption prediction model is a comparatively effective and efficient model for prediction using data obtained from real-driving measurements and is suitable for real-world applications for large data sets.

Fig. 4(a), (b) and (c) show CO_2, NO_x and fuel consumption results for all three models. Fig. 4(a) shows the approximation degree between the CO_2 prediction results (LR, SVM and GPR) and the actual CO_2. The figure shows that some points have a large deviation in the prediction results of the LR and SVM models. This is also evident from their model performance indicators. Overall, the three models have a good fitting degree as the predicted results are generally distributed on both sides of $y = x$ with a high approximation degree. NO_x and fuel consumption prediction results presented in Fig. 4(b) and (c) also show a similar distribution with a high approximation degree. Moreover, the best model performance is shown in Table 4 for NO_x and fuel consumption also showed the best fitting degree in Fig. 4(b) and (c) respectively.

Feature importance sore has been calculated for the developed models to assess the influence of demographic variables on emission and fuel consumption. Fig. 5 presents the impact of four demographic variables namely, driver age, driving experience, driver gender and traffic condition on CO_2, NO_x emission and fuel consumption. Fig. 5 shows driver age has a high impact on NO_x emission and driving experience has a high impact on fuel consumption. On the other hand, driver gender has very little impact on CO_2, NO_x emissions and fuel consumption; the impact on NO_x is almost negligible. Road environment i.e., traffic conditions have a strong impact on CO_2, NO_x emissions and fuel consumption, among these, CO_2 showed the strongest impact.

The Pearson correlation coefficient presented in Table 4 showed driving behaviour and road environment-related parameters such as acceleration, deceleration, idling, speed, RPA, VA and RCS have a strong correlation with CO_2 and NO_x emissions, hence, used in the prediction models. Several real-driving measurement studies also reported driving behaviour and traffic condition to have a strong influence on emissions [14,22,34,36,53]. For example, André and Rapone, [2] studied RDE in various traffic environments and observed congested traffic is responsible for most of the acceleration and NOx emission. Varella et al., [53] in their RDE study observed a 55% increment in NO_x emissions and a 7% increment in CO_2 emissions for variation in driving style (normal to aggressive).

RPA is a popular metric for driving style characterisation [10,34]. According to the RDE regulation guidelines, during urban driving RPA ≤ 0.13 m/s^2 refers to timid driving, > 0.13 m/s^2 and < 0.2 m/s^2 refers to normal driving and ≥ 0.2 m/s^2 refers to aggressive driving [18]. Among 60 performed trips, the current study has found that 43% of the trips were aggressive. An increased dynamicity often occurs from a speed adaptation tendency after stopping the vehicle at traffic signals and slowing to navigate traffic features (roundabouts/sharp corners) during urban driving. Congested traffic during busy hours causes subsequent aggressive driving. Moreover, demographic variables such as driver age and experience also have an impact on emissions as shown in Fig. 5. Setup control in driving behaviour by implementing an artificially relaxed behaviour i.e., eco-driving could play an important role to reduce the driving dynamic values near the limit imposed by RDE regulations, therefore, contributing to pollutant emission reduction. Current eco-driving practices mainly focus on fuel consumption and CO_2 reduction but ignore other pollutant emissions [3,5,27,28]. Therefore, incorporating driver behaviour into eco-driving strategies and establishing driver education protocol considering these factors can significantly contribute to pollutant emission reduction, as well as reducing fuel consumption. Such training programs will provide drivers with the knowledge (theoretical training) and skills (practical training) to drive in a more fuel-efficient way, as well as being able to reduce pollutant emissions.

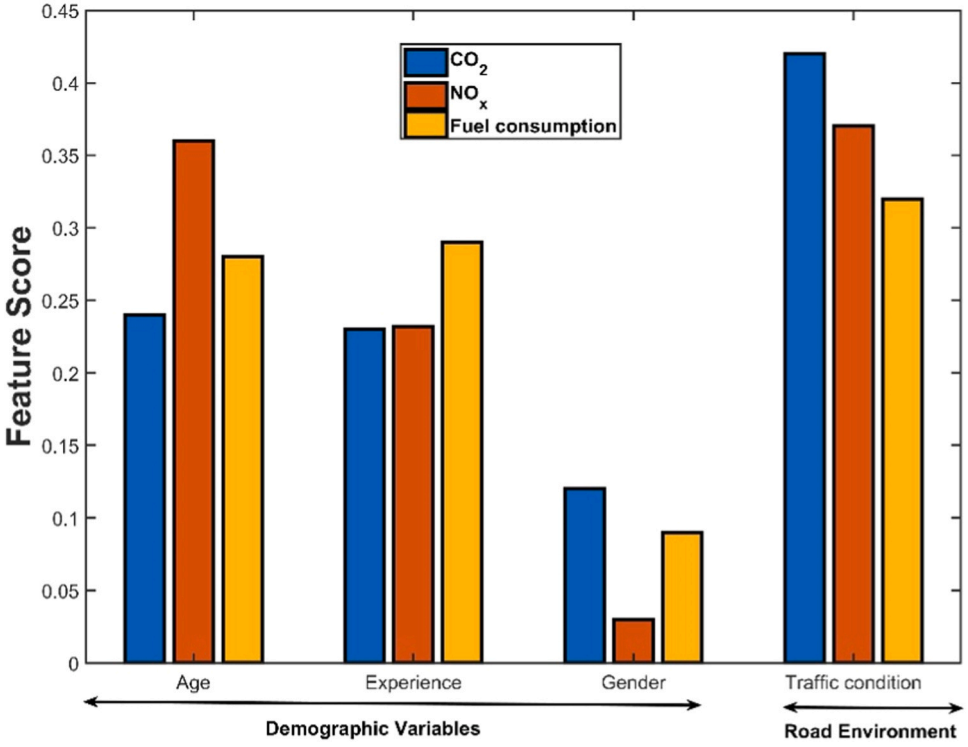

Fig. 5. Impact of demographic variables and road environment on CO_2, NO_x emissions and fuel consumption.

5. Conclusion

In this study, instantaneous vehicle movement information, engine running information, real-time raw exhaust emissions and fuel consumption were recorded and analysed to characterise the pattern for modelling using a PEMS for an urban transient driving environment. Machine learning approaches were used to estimate CO_2, NO_x and fuel consumption using the data obtained through real-driving measurements. The correlation between the recorded parameters and target parameters (CO_2, NO_x and fuel consumption) was analysed and the relevant input parameters for each target parameter were extracted through the filter-based feature selection method. Using the selected predictor variables associated with each target variable three different prediction models have been developed (LR, SVM and GPR).

GPR and SVM-based models presented better predictions of CO_2 emissions with a lower RMSE and ARE and a higher R^2 range. Among these, GPR is the most suitable model for CO_2 prediction as LR and SVM show 2.45 and 2.06 times higher ARE respectively. The lower ARE for GPR indicates that the predicted results are very close to the measurements.

The LR-based model shows a significantly higher fitting degree for NO_x emission compared to SVM and GPR. A comparatively higher ARE than CO_2 and fuel consumption indicates the uncertainty associated with NO_x emission in the real-driving scenario.

All three models predicted fuel consumption with higher accuracy. Among these, SVM showed the highest fitting degree and the lowest RMSE and ARE.

This study developed machine learning-based emission and fuel consumption estimation models using data obtained from a large number of real-world drivers than existing studies. The PEMS sensors were installed before the after-treatment system, rather than at the exit of the tailpipe as in conventional methods, to capture raw emission data that is strongly correlated with driving dynamics. The result of this study indicates a significant improvement in the accuracy of estimating on-road emissions and fuel consumption during highly transient driving environments. This study will aid to obtain more realistic emission and fuel consumption estimations that reflect the real-world scenario. Extending these modelling approaches to other emissions (for example, CO, HC and particulate matter) and different vehicle classes, traffic conditions, and road and fuel types is recommended.

The insights of this study will contribute to developing emission-optimal driving assistance systems that will reduce the environmental impact as well as the amount of fuel burned. Moreover, the findings of this study could be useful for developing driver education programs to minimise aggressive driving behaviour that is responsible for detrimental pollutant emissions in urban areas.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to acknowledge all the participant drivers for their significant contribution to the emission measurement campaign. We also acknowledge the Space, Assets, and Logistics team (QUT, Science and Engineering Faculty) for their support in facilitating access to the test vehicle. Dr Amir Moghaddam is acknowledged for his continuous support during the experiments.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.nxener.2023.100060.

References

[1] A. Alin, Multicollinearity, Wiley Interdiscip. Rev. Comput. Stat. 2 (2010) 370–374.
[2] M. André, M. Rapone, Analysis and modelling of the pollutant emissions from European cars regarding the driving characteristics and test cycles, Atmos. Environ. 43 (2009) 986–995.
[3] K. Ayyildiz, F. Cavallaro, S. Nocera, R. Willenbrock, Reducing fuel consumption and carbon emissions through eco-drive training, Transp. Res. Part F. Traffic Psychol. Behav. 46 (2017) 96–110.
[4] Bakhshmand, S.K., Mulholland, E., Tietge, U., Rodríguez, F., 2022. Remote sensing of heavy-duty vehicle emissions in Europe, (No. 2022–25).
[5] D. Barić, G. Zovak, M. Perišа, Effects of eco-drive education on the reduction of fuel consumption and CO_2 emissions, Promet-Traffic&Transp. 25 (2013) 265–272.
[6] Barlow, T.J., Latham, S., McCrae, I.S., Boulter, P.G., 2009. A reference book of driving cycles for use in the measurement of road vehicle emissions. Version 3, TRL Published Project Report PPR354.
[7] S. Baumann, U. Klingauf, Modeling of aircraft fuel consumption using machine learning algorithms, CEAS Aeronaut. J. 11 (2020) 277–287.
[8] D.A. Belsley, E. Kuh, R.E. Welsch, Regression diagnostics: identifying influential data and sources of collinearity, Wiley Ser. Probab. Math. Stat. (1980) 571.
[9] R. Berk, An introduction to statistical learning from a regression perspective, in: Handbook of Quantitative Criminology, Springer, 2010, pp. 725–740.
[10] T.A. Bodisco, S.M.A. Rahman, F.M. Hossain, R.J. Brown, On-road NOx emissions of a modern commercial light-duty diesel vehicle using a blend of tyre oil and diesel, Energy Rep. 5 (2019) 349–356.
[11] N. Brusselaers, C. Macharis, K. Mommens, Rerouting urban construction transport flows to avoid air pollution hotspots, Transp. Res. Part D. Transp. Environ. 119 (2023), 103747.
[12] Y. Cheng, L. He, W. He, P. Zhao, P. Wang, J. Zhao, K. Zhang, S. Zhang, Evaluating on-board sensing-based nitrogen oxides (NOx) emissions from a heavy-duty diesel truck in China, Atmos. Environ. 216 (2019), 116908.
[13] Dabbas, W.M., 2010. Modelling vehicle emissions from an urban air-quality perspective: testing vehicle emissions interdependencies.
[14] I. De Vlieger, D. De Keukeleere, J.G. Kretzschmar, Environmental effects of driving behaviour and congestion related to passenger cars, Atmos. Environ. 34 (2000) 4649–4655.
[15] Donateo, T., Giovinazzi, M., 2018. Some repeatability and reproducibility issues in real driving emission tests. SAE Technical Paper.
[16] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Adv. Neural Inf. Process. Syst. 9 (1997) 155–161.
[17] M.R. Edwards, M.M. Klemun, H.C. Kim, T.J. Wallington, S.L. Winkler, M.A. Tamor, J.E. Trancik, Vehicle emissions of short-lived and long-lived climate forcers: trends and tradeoffs, Faraday Discuss. 200 (2017) 453–474.
[18] European Commission, 2016. Commission Regulation (EU) 2016/427 of 10 March 2016 Amending Regulation (EC) No 692/2008 as Regards Emissions from Light Passenger and Commercial Vehicles (Euro 6).
[19] X. Fang, N. Papaioannou, F. Leach, M.H. Davy, On the application of artificial neural networks for the prediction of NO x emissions from a high-speed direct injection diesel engine, Int. J. Engine Res. 22 (2021) 1808–1824.
[20] G. Fontaras, V. Franco, P. Dilara, G. Martini, U. Manfredi, Development and review of Euro 5 passenger car emission factors based on experimental results over various driving cycles, Sci. Total Environ. 468 (2014) 1034–1042.
[21] J.C. Fussell, M. Franklin, D.C. Green, M. Gustafsson, R.M. Harrison, W. Hicks, F. J. Kelly, F. Kishta, M.R. Miller, I.S. Mudway, A review of road traffic-derived non-exhaust particles: emissions, physicochemical characteristics, health risks, and mitigation measures, Environ. Sci. Technol. 56 (2022) 6813–6835.
[22] J. Gallus, U. Kirchner, R. Vogt, T. Benter, Impact of driving style and road grade on gaseous exhaust emissions of passenger vehicles measured by a Portable Emission Measurement System (PEMS), Transp. Res. Part D. Transp. Environ. 52 (2017) 215–226.
[23] B. Giechaskiel, R. Suarez-Bertoa, T. Lähde, M. Clairotte, M. Carriero, P. Bonnel, M. Maggiore, Evaluation of NOx emissions of a retrofitted Euro 5 passenger car for the Horizon prize "Engine retrofit, Environ. Res. 166 (2018) 298–309.
[24] C. Guo, B. Yang, O. Andersen, C.S. Jensen, K. Torp, Ecomark 2.0: empowering eco-routing with vehicular environmental models and actual vehicle fuel consumption data, Geoinformatica 19 (2015) 567–599.
[25] N. Hashemi, N.N. Clark, Artificial neural network as a predictive tool for emissions from heavy-duty diesel vehicles in Southern California, Int. J. Engine Res. 8 (2007) 321–336.
[26] Z. Hu, Y. Jin, Q. Hu, S. Sen, T. Zhou, M.T. Osman, Prediction of fuel consumption for enroute ship based on machine learning, IEEE Access 7 (2019) 119497–119505.
[27] Y. Huang, E.C.Y. Ng, J.L. Zhou, N.C. Surawski, E.F.C. Chan, G. Hong, Eco-driving technology for sustainable road transport: a review, Renew. Sustain. Energy Rev. 93 (2018) 596–609.
[28] I. Jeffreys, G. Graves, M. Roth, Evaluation of eco-driving training for vehicle fuel use and emission reduction: A case study in Australia, Transp. Res. Part D. Transp. Environ. 60 (2018) 85–91.
[29] M.M.R. Komol, M.M. Hasan, M. Elhenawy, S. Yasmin, M. Masoud, A. Rakotonirainy, Crash severity analysis of vulnerable road users using machine learning, PLoS One 16 (2021), e0255828.
[30] S. Kwon, Y. Park, J. Park, J. Kim, K.-H. Choi, J.-S. Cha, Characteristics of on-road NOx emissions from Euro 6 light-duty diesel vehicles using a portable emissions measurement system, Sci. Total Environ. 576 (2017) 70–77.
[31] C.M.A. Le Cornec, N. Molden, M. van Reeuwijk, M.E.J. Stettler, Modelling of instantaneous emissions from diesel vehicles with a special focus on NOx: Insights from machine learning techniques, Sci. Total Environ. 737 (2020), 139625.

[32] H. Le Thi, Health Impacts of Traffic-related Air Pollution: Cause-effect Relationships and Mitigating Measures, in: CIGOS 2019, Innovation for Sustainable Infrastructure: Proceedings of the 5th International Conference on Geotechnics, Civil Engineering Works and Structures, Springer, 2020, pp. 1031–1036.
[33] J.M. Luján, V. Bermúdez, V. Dolz, J. Monsalve-Serrano, An assessment of the real-world driving gaseous emissions from a Euro 6 light-duty diesel vehicle using a portable emissions measurement system (PEMS), Atmos. Environ. 174 (2018) 112–121.
[34] J.M. Luján, C. Guardiola, B. Pla, V. Pandey, Impact of driving dynamics in RDE test on NO x emissions dispersion, Proc. Inst. Mech. Eng. Part D. J. Automob. Eng. 234 (2020) 1770–1778.
[35] J. May, D. Bosteels, C. Favre, An assessment of emissions from light-duty vehicles using PEMS and chassis dynamometer testing, SAE Int. J. Engines 7 (2014) 1326–1335.
[36] Z. Mera, N. Fonseca, J.-M. López, J. Casanova, Analysis of the high instantaneous NOx emissions from Euro 6 diesel passenger cars under real driving conditions, Appl. Energy 242 (2019) 1074–1089.
[37] E. Moradi, L. Miranda-Moreno, Vehicular fuel consumption estimation using real-world measures through cascaded machine learning modeling, Transp. Res. Part D. Transp. Environ. 88 (2020), 102576.
[38] M.H. Moradi, A. Heinz, U. Wagner, T. Koch, Modeling the emissions of a gasoline engine during high-transient operation using machine learning approaches, Int. J. Engine Res. (2021), 14680874211032380.
[39] M.H. Moradi, A. Sohani, M. Zabihigivi, H. Wirbser, A comprehensive approach to find the performance map of a heat pump using experiment and soft computing methods, Energy Convers. Manag. 153 (2017) 224–242.
[40] Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. Applied linear statistical models.
[41] M.J. Nieuwenhuijsen, Urban and transport planning pathways to carbon neutral, liveable and healthy cities; A review of the current evidence, Environ. Int. 140 (2020), 105661.
[42] L. Pelkmans, D. De Keukeleere, H. Bruneel, G. Lenaers, Influence of vehicle test cycle characteristics on fuel consumption and emissions of city buses, SAE Trans. 1388 (2001) 1398.
[43] F. Perrotta, T. Parry, L.C. Neves, Application of machine learning for fuel consumption modelling of trucks, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 3810–3815.
[44] S. Prakash, T.A. Bodisco, An investigation into the effect of road gradient and driving style on NOx emissions from a diesel vehicle driven on urban roads, Transp. Res. Part D. Transp. Environ. 72 (2019) 220–231.
[45] L. Qu, M. Li, D. Chen, K. Lu, T. Jin, X. Xu, Multivariate analysis between driving condition and vehicle emission for light duty gasoline vehicles during rush hours, Atmos. Environ. 110 (2015) 103–110.
[46] Regulation, C., 2016. Regulation (EC) No 692/2008 as regards emissions from light passenger and commercial vehicles (Euro 6),". Eur. Union.
[47] M.A. Schroeder, J. Lander, S. Levine-Silverman, Diagnosing and dealing with multicollinearity, West. J. Nurs. Res. 12 (1990) 175–187.
[48] G.M.H. Shahariar, T.A. Bodisco, A. Zare, M. Sajjad, M.I. Jahirul, T.C. Van, H. Bartlett, Z. Ristovski, R.J. Brown, Impact of driving style and traffic condition on emissions and fuel consumption during real-world transient operation, Fuel 319 (2022), 123874.
[49] G.M.H. Shahariar, M. Sajjad, K.A. Suara, M.I. Jahirul, T. Chu-Van, Z. Ristovski, R. J. Brown, T.A. Bodisco, On-road CO_2 and NOx emissions of a diesel vehicle in urban traffic, Transp. Res. Part D. Transp. Environ. 107 (2022), 103326.
[50] M. SjÖgren, H. Li, U. Rannug, R. Westerholm, Multivariate analysis of exhaust emissions from heavy-duty diesel fuels, Environ. Sci. Technol. 30 (1995) 38–49.
[51] R. Suarez-Bertoa, V. Valverde, M. Clairotte, J. Pavlovic, B. Giechaskiel, V. Franco, Z. Kregar, C. Astorga, On-road emissions of passenger cars beyond the boundary conditions of the real-driving emissions test, Environ. Res. 176 (2019), 108572.
[52] C.G. Thompson, R.S. Kim, A.M. Aloe, B.J. Becker, Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results, Basic Appl. Soc. Psych. 39 (2017) 81–90.
[53] R.A. Varella, M.V. Faria, P. Mendoza-Villafuerte, P.C. Baptista, L. Sousa, G. O. Duarte, Assessing the influence of boundary conditions, driving behavior and data analysis methods on real driving CO_2 and NOx emissions, Sci. Total Environ. 658 (2019) 879–894.
[54] G.J.M. Velders, G.P. Geilenkirchen, R. de Lange, Higher than expected NOx emission from trucks may affect attainability of NO_2 limit values in the Netherlands, Atmos. Environ. 45 (2011) 3025–3033.
[55] M. Weiss, P. Bonnel, R. Hummel, A. Provenza, U. Manfredi, On-road emissions of light-duty vehicles in Europe, Environ. Sci. Technol. 45 (2011) 8575–8581.
[56] Y. Yao, X. Zhao, C. Liu, J. Rong, Y. Zhang, Z. Dong, Y. Su, Vehicle fuel consumption prediction method based on driving behavior data collected from smartphones, J. Adv. Transp. 2020 (2020).
[57] J. Yuan, V. Nian, Ship energy consumption prediction with Gaussian process metamodel, Energy Procedia 152 (2018) 655–660.
[58] W. Zeng, T. Miwa, T. Morikawa, Exploring trip fuel consumption by machine learning from GPS and CAN bus data, J. East. Asia Soc. Transp. Stud. 11 (2015) 906–921.