Perrotta, Federico and Parry, Tony and Neves, Luís C. (2017) Application of machine learning for fuel consumption modelling of trucks. In: 2017 IEEE International Conference on Big Data, 11-14 Dec 2017, Boston, USA.

Application of machine learning techniques to estimate truck fleet fuel consumption from telematic data and road geometry

Federico Perrotta[a], Tony Parry[a], Luis Canhoto Neves[b], Mohammad Mesgarpour[c]

[a] Nottingham Transportation Engineering Centre, Faculty of Engineering, University of Nottingham, University Park, Nottingham NG7 2RD, UK

[b] Resilience Engineering Research Group, Faculty of Engineering, University of Nottingham, University Park, Nottingham NG7 2RD, UK

[c] Microlise Ltd, Farrington Way, Eastwood, Nottingham NG16 3AG, UK

Abstract

This paper presents a comparison of the performance of three machine learning techniques in estimating the fuel consumption of trucks using telematic data and road geometry information. A large amount of data is collected from sensors installed on trucks according to standard SAE J1939. They are used to constantly monitor the performance of the vehicles by fleet managers and inform their decisions regarding training of drivers and truck maintenance. The data used here describe the performance of 1,110 articulated trucks driving during one week, on two motorways in England. From the analysis of the Pearson's

correlation coefficients, p-values, adjusted-R^2, and Lasso, the key variables influencing fuel consumption were identified. From this, four models have been generated and their performance compared. These are a multiple linear regression, and three machine learning models; a Support Vector Machine (SVM), a Random Forest (RF), and an Artificial Neural Network (ANN). The paper shows how machine learning techniques can significantly improve the accuracy of predictions compared to the linear regression model, reducing variance in the final estimates. Finally, a parametric analysis was performed to estimate the impact of each of the selected variables on the fuel consumption of the fleet of trucks considered.

Keywords: fuel consumption, big data, machine learning, truck fleet management, road asset management, telematic data

1. Introduction

In 2015, the transport sector, consisting of road transport, railways, aviation, and shipping, was responsible for close to a quarter of greenhouse gas (GHG) emissions in the UK [1]. Of this, road transport was the most significant source of emissions. Although since the early 2000s the emissions per vehicle have generally decreased, the overall fuel consumption has increased, due to increased vehicle kilometers traveled [1]. A better understanding of the impact on fuel consumption of factors external to the road vehicles, such as road geometry, may be important in developing strategies to reduce GHG emissions.

Following international regulations, standard tests are performed to measure GHG emissions and fuel consumption of road vehicles (e.g. the European NEDC, the American FTP-75, the Japanese JC08, or the internationally harmonized WLTP, WHDC, etc.). However, these measure test vehicles only under standard drive cycles, which may not reflect what happens in reality. For this reason, many experts (e.g. [2–7]) criticize the regulations, arguing that the

unrealistic test speeds, lack of hill climbing included in the test cycle etc. may not properly represent the conditions of real drive cycles.

Based on these measurements a variety of fuel consumption models are available in the literature (e.g. [8–13]), however, most of these models are calibrated for light vehicles or offer only simplified mathematical or physical-mechanistic expressions to compute the instantaneous fuel consumption of the vehicles considered, making assumptions about the driving mode and without considering the whole performance of each vehicle or the characteristics of the road. Therefore, the results of these models may also not be representative of what happens in real driving conditions.

A new fuel consumption model based on real data from the actual road network can help engineers in addressing those factors affecting fuel consumption and related GHG emissions from the road transport industry. For example, this may be useful for highway asset managers in their decision making process regarding the geometric design of new roads or, for truck fleet managers in making decisions about the routing of their vehicles. Nowadays, GPS systems select routes based on the duration of travel, traffic, presence of accidents or road works, length of the route and highway tolls, among other factors (Fawcett & Robinson 2000). Considering fuel economy and in particular, the impact of road geometry on fuel consumption, may help fleet managers in reducing fuel costs and GHG emissions.

Fuel consumption of road vehicles is a complex problem determined by a number of variables e.g. the type of vehicle, the type of engine, the gross vehicle weight, the road gradient, the wind speed and its direction, etc. [14,15]. From a physical point of view, energy losses can be attributed to four factors 1) aerodynamics, caused by the friction generated by the surrounding air on the vehicle surface, 2) the rolling resistance, depending mainly on the vehicle speed, the tire pressure and the road surface conditions, 3) the gradient of the road, and 4) the internal friction, due to the inertia of the vehicle, the rotating masses in the

powertrain and the brakes (Guzzella & Sciarretta 2015). Therefore, using a physical based approach, fuel consumption has been modelled by using Equation (1):

$$m\frac{d}{dt}v(t) = F_t(t) - (F_a(t) + F_r(t) + F_g(t) + F_d(t)) \quad (1)$$

where m is the mass of the vehicle, $\frac{d}{dt}v(t)$ is the partial derivative of the vehicle speed with respect to time t, F_t is the traction force, F_r is the rolling resistance, F_g the gravitational force, and F_d a 'disturbance force' summarizing all other effects not specified. However, due to the complexity of the system and the dependence of the considered variables on v and t, Eq. (1) is usually applied to the analysis of standard drive cycles performed under well-known conditions and by considering very short time windows, to make the partial derivative of v constant in respect to t. This requires the collection of high frequency data, high computational power and calculation time and suffers from limitations in the number of cases that can be considered and the validity of the results obtained due to the assumptions made in the model (e.g. constant speed, acceleration, road gradient, a short time window, on a short road segment, etc.). Therefore, results obtained by applying this approach may be far from what happens in real driving conditions and the model must be recalibrated, and new experiments performed, each time a new vehicle is developed. Examples of models that use this approach to estimate the fuel consumption of road vehicles are the HDM-4, Highway Development and Maintenance model [16], used worldwide in road asset management for the estimation of the impact of the road infrastructure conditions on vehicle fuel economy and vehicle operating costs, and PERE, the Physical Emission Rate Estimator [17], used internationally to estimate the emission rates of conventional and advanced technology vehicles.

Another approach for estimating the fuel consumption of road vehicles and addressing its causes is the use of regression techniques. In the past, multiple methods have been used, but what characterizes this approach is its speed, and several applications have demonstrated the

ability to handle large quantities of data, giving relatively precise estimates in a short time. For this reason, multiple models can be derived for different situations and they can include non-standard conditions. Examples of studies that used regression techniques to analyse fuel consumption are the works of Clark et al. [9], Lee et al. [18] and Zeng et al. [19], however, none of these models have yet found wide application in practice at international level.

Machine learning regression algorithms are commonly applied to find trends, predict future performance and identify relationships in a range of subjects including computer vision (e.g. [20]), health data monitoring (e.g. [21]), bank fraud detection (e.g. [22]), etc. Regarding the prediction of vehicle fuel consumption these advanced algorithms have been applied to aircraft using artificial neural networks (ANNs) [23,24] and, more recently, to HGVs and buses using support vector machine (SVM), random forests (RFs), and ANNs [8,18,25,26]. However, all the studies used a limited number of vehicles tested under carefully controlled conditions on a few selected road segments with known geometry. Consequently, the applicability of machine learning to data gathered under real driving conditions, including aspects such as weather conditions, or interactions with other traffic, is yet to be evaluated.

This paper aims to analyse a large dataset of fuel consumption records for conditions representative of vehicles driving on motorways in the UK. The data used is collected using on board sensors fitted as standard [27] and transmitted telematically. A total of 14,281 records from 1,110 articulated trucks driving at relatively constant speed on 300km of motorway have been considered. Four regression models have been developed and their performance compared. These are a multiple linear regression, a Support Vector Machine (SVM), a Random Forest (RF), and an Artificial Neural Network (ANN). The generated models estimate the fuel consumption of the considered fleet of trucks expressed as litres per 100km (l/100km). The three main stages of the analysis are:

i. selection of the most significant variables is performed by comparing the results of multiple statistics including the p-values, the adjusted-R^2, and the Lasso regression;
ii. the four models, linear regression, SVM, RF, and ANN are generated; ten-fold cross-validation was performed to validate the machine learning models;
iii. a parametric analysis is performed to evaluate the impact of each of the variables included in the model on the fuel consumption of the considered fleet of trucks.

The objectives of the study can be summarised as follows:

- define a new methodology for the estimation of fuel consumption based on vehicle telematic data and road geometry information;
- compare the performance of multiple linear regression with machine learning methods, SVM, RF, and ANN, in predicting fuel consumption of a large fleet of trucks;
- test the capability of machine learning methods to address and estimate the impact of each variable included in the generated model by performing a parametric analysis;
- improve knowledge about the estimation of fuel consumption by analysing real data from an actual road network.

The generated models are useful for truck fleet managers for re-routing of vehicles, by predicting the possible difference in fuel consumption on alternative roads, and for road asset managers who can use the models to estimate possible savings and reduction of GHG emissions for new roads by changing the geometry of the infrastructure (i.e. road gradient and curvature).

2. Data

2.1 Truck telematic data

Modern trucks are equipped with sensors according to standard SAE J1939 [27]. These continuously measure multiple parameters helping fleet managers in monitoring the performance of their trucks and in taking decisions regarding vehicle maintenance and driver training. For this study, anonymised data were provided by Microlise Ltd. No information about the driver, the maker, or the company owning the vehicles is included in the analysed data. The vehicle speed, the percent torque, the revolutions per minute (rpm) of the engine, the use of cruise control, the use of brakes and acceleration pedal, the traveled distance and the fuel used are available, among many other parameters. An estimate of gross vehicle weight is also reported, as calculated by an algorithm based on speed and engine data, such as torque and rpm. Cumulative fuel consumption is recorded, for the selected trucks to the nearest 0.001 litres. The data are georeferenced with the geographical position of all vehicles constantly monitored through GPS. This information is collected as default each 1 minute or 1 mile (approximately 1.6 km) travelled or when triggered by other events, such as change in gear etc., and is downloaded telematically. Each vehicle can be identified only by an ID number in reference to the electronic control unit (ECU) installed on the vehicle, the wheel configuration, and the type of engine. Date and time, the geographical position of the vehicle when the event is triggered, and an unambiguous ID number identifies each of the recorded events. This information allows the whole journey of each vehicle to be retraced.

2.2 Road geometry and condition data

Every year Highways England and its partners update the information available regarding the strategic road network in England. This includes, for example, construction details, condition of the road pavement, etc. and is stored in the Highways Agency Pavement Management

System (HAPMS). These data are collected mostly for quality control and for strategic decision making regarding maintenance and rehabilitation (M&R) of the infrastructure. The data are collected automatically using laser sensors installed on a monitoring vehicle. For instance, the vehicle measures the road gradient, crossfall, wheelpath rutting, presence of potholes, the roughness of the pavement at different wavelengths, the surface macrotexture and the radius of curvature, etc. These are georeferenced (through GPS) and usually reported each 10m and stored for each 100m average. Each record refers to a specific road section and direction of travel.

2.3 Data mining

In this case study, in order to simplify the data analysis and reduce the effect of nonlinearity on the fuel consumption of the considered fleet of trucks, only journeys performed by a single type of truck at a constant speed (+/- 2.5km/h) along the selected road segments have been considered. Therefore, from the 594,690 records initially available, during a week in October 2016 along two segments of motorway M1 and the entire M18, only 14,281 records from 1,110 Euro 6 articulated trucks are considered in the study. These trucks were selected for the reporting precision of their fuel measurement devices (earlier trucks typically report only to the nearest 0.1l). The total length of the considered road segments is 300km. The selected records are for 1 minute or 1 mile of travel (whichever is the shorter) and records triggered by other events were discarded. Table 1 reports all the available measurements with a brief qualitative description and their resolution. Each record includes measurements of the instantaneous vehicle speed, date and time, and distance travelled by the vehicle. Each record is considered to be performed at a constant speed and selected for analysis, if the absolute difference between the instantaneous speed measured in consecutive records does not differ by more than 2.50km/h, and, the absolute difference between the average of the speeds at the initial and final records do not differ by more than 2.50km/h.

Table 1 - Summary of all the measurements available in the database and their resolution.

Variable name	Description	Resolution
Vehicle.ID	Vehicle identification code	-
EventID.Start	ID code assigned to the initial event	-
EventID.End	ID code assigned to the final event	-
Wheel.Plan	Nr. of wheels on the tractor + trailer	-
Euro.Type	Type of engine	-
Date.Start	Date/time of initial location	'dd/MM/yyyy hh:mm:ss'
Date.End	Date/time of final location	'dd/MM/yyyy hh:mm:ss'
Lat.Start	Latitude of the initial location	0.0001º
Long.Start	Longitude of the initial location	0.0001º
Lat.End	Latitude of the final location	0.0001º
Long.End	Longitude of the final location	0.0001º
Direction.Start	Heading of the vehicle from the North	1.00º
Direction.End	Heading of the vehicle from the North	1.00º
Altitude.Start	Altitude at initial location	1.00 m
Altitude.End	Altitude at final location	1.00 m
Travelled.Dist	Distance between two considered locations	0.01 m
Travelled.Time	Time between two considered locations	0.1 s
Road.ID	ID code of the considered road	-
Gross.Weight_kg	Estimated gross vehicle weight (GVW)	400 kg
Speed.Start	Instantaneous speed at the initial location	0.01 km/h
Speed.End	Instantaneous speed at the final location	0.01 km/h
Acceleration	Average acceleration	0.01 m/s^2
Speed.AVG	Average speed	0.01 km/h
Travel.Direction	Average direction of travel	1.00º
Gear.Start	Gear used at the initial location	-
Gear.End	Gear used at the final location	-
CruiseControl.Start	1/0 message for cc activation, initial location	-
CruiseControl.End	1/0 message for cc activation, final location	-
Torque.Start	Torque %, initial location	1.00 %
Torque.End	Torqe %, final location	1.00 %
Torque.AVG	Mean of Torque.Start and Torque.End	1.00 %
Revs.Start	Instantaneous rpm at the initial location	1.00 rpm
Revs.End	Instantaneous rpm at the final location	1.00 rpm
Revs.AVG	Mean of Revs.Start and Revs.End	1.00 rpm
Used.Fuel	Quantity of fuel used in the journey	0.001 l
Fuel.Consumption	The ratio between the Used.Fuel and Travelled.Distance in l/100km	0.01 l/100km
Const.Speed	1/0 message if speed is constant	-
Geom. Radius	Calculated as the mean of the absolute value of radius of curvature	1.00 m
Geom. Radius_sd	The standard deviation of Geom. Radius	-
Geom.Gradient	Calculated as the mean of the road gradient	0.01 %
Geom.Gradient_sd	The standard deviation of Geom.Gradient	-
Geom.Crossfall	Calculated as the mean of the crossfall	0.01 %
Geom.Crossfall_sd	The standard deviation of Geom.Crossfall	-

Also, in order to exclude from the analysis data from vehicles with technical issues and short time intervals, journeys shorter than 150m have been excluded. Fuel consumption lower than 3.90 l/100km or higher than 60 l/100km are very rare in the remaining data (< 0.5% of the data) and are excluded as outliers caused by unrecorded events or technical problems in the engine or in the data collection system. After applying these filters to the data, 14,281 records are available in total. Road geometry records were combined with truck data records by comparing their geographical position.

3. Methodology

The software used for analysing the data is R v. 3.4.1 (CRAN 2017). The major packages used to perform the analysis are the '`glmnet`' [28] and the '`caret`' [29] used to perform the Lasso regression and the ten-fold cross-validation of the generated models respectively, the '`e1070`' [30] used to build the SVM model, the '`randomForest`' [31] used to build the RF model, and the '`neuralnet`' [32] used to build the ANN model.

3.1 Variables selection

In order to avoid overfitting, among all the parameters available in the database only the most significant variables have been selected and included in the regression analysis. Five different statistics were used to make the selection: 1) the *Pearson's correlation coefficient* (r), 2) the *p-values*, 3) the *adjusted-R^2*, 4) a *Lasso regression* and 5) Random Forests. In particular:

1) An initial cut-off of the variables was performed by excluding all the variables that have a poor correlation (lower than 0.10) with fuel consumption;
2) At this point, p-values for all remaining variables were computed to test the significance of each. Only variables with a p-value < 0.05 are considered to be statistically significant and therefore included in the regression analysis;

3) Then, the adjusted-R^2 was computed for each possible model (with different combinations of variables) and the one showing the highest adjusted-R^2 was selected;
4) In order to test the significance of the selected parameters (avoiding overfitting), a Lasso regression [33] was applied to the variables.

   'Lasso' stands for Least Absolute Shrinkage and Selection Operator. It is a shrinkage method that is generally used for regression but that can also be used for variable selection. It is similar to a linear regression in which parameters are estimated with the least square method, however, in the case of Lasso, the regression coefficients are not computed by minimizing only the residual sum of squares (RSS) but the quantity:

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \quad (2)$$

   where, β_j is the regression coefficient associated with the variable *i*, and λ is a tuning parameter. The Lasso differs from the least square method because of the term $\lambda \sum_j |\beta_j|$. This is called the 'shrinkage penalty' and its effect is to reduce the estimates of the β_j parameters towards zero. This term helps in reducing variance significantly, aiming to improve the overall fit of the regression by excluding overfitting variables [34]. It is possible to select the best value for λ by introducing multiple values for the tuning parameter and selecting the one that reduces the mean square error of the Lasso to a minimum. Only the variables with coefficients that are not reduced to zero by the Lasso are adopted in the regression analysis. Due to its properties, Lasso can be used to perform variable selection [34] with the possibility of also detecting nonlinear correlations between the variables (e.g. [35]). Therefore, Lasso can also be used to test the significance of Boolean variables such as the activation of the cruise control in this study.

5) At this point, the Random Forest algorithm is used to check the findings obtained by looking at the p-values, adjusted-R^2, and the Lasso. Random Forests can detect nonlinear relationships and deal with discrete variables being able to identify complex but significant correlations [36]. In particular the algorithm calculates the 'increase in node purity' that each variable brings to the model. Higher 'increase in node purity' means that including that variable helps in reducing the variance that the model is not able to explain [36]. Further details about how this algorithm works is given later in its dedicated sub-chapter.

3.2 Linear regression

As a first step, a linear regression of the selected variables was fitted to the data and used to analyze the fuel consumption of the considered fleet of trucks. Linear regression has previously been used to make predictions of fuel consumption of road vehicles based on drive cycle properties (e.g. [9,10,12]). For this reason, this study uses the generated linear regression model as a reference for judging the performance of the applied machine learning regression techniques.

3.3 Machine learning

The main reason to apply machine learning to this specific case study is the ability of these techniques to learn from data, recognising specific patterns and complex relationships, making predictions based on them. The user does not have to specify the type of relation between the variables which is determined by the algorithm. Also, thanks to their capability of dealing with large quantities of data, machine learning techniques are applied nowadays to a number of different areas in academia and in industry being considered as the most advanced tools for solving any sort of classification and regression.

In order to use machine learning regression techniques, one has first to train the model on existing data and then test it in a new situation using new data. In practice, the available dataset is usually split in two and 75% of data is used for training and the remaining 25% is used to test the generated model. The split of data must be made randomly in order to reduce bias in the final estimates.

3.3.1 Support Vector Machine

Support Vector Machine (SVM) [37,38] is a machine learning discriminative classifier algorithm characterized by the ability to control the decision function by defining a kernel function that identifies one or multiple separating hyperplanes. Nowadays, although the mathematics behind SVM is complex [39,38], this method is widely used in practical applications (e.g. [40,41]). In the past, to solve similar problems concerning fuel consumption of road vehicles, the radial basis function (RBF) has been selected as the kernel function for regression and has been selected in this study to develop the SVM model. This is because the RBF maps samples into a higher dimensional space and can handle the case when the relation between class labels and attributes is nonlinear. The grid-search method has been used to determine the optimal parameters to use in the model. For this study, the SVM model has been developed using the e1071 R package [30]. This provides an interface to the `libsvm` C++ library [42] and is a powerful toolkit for SVM application.

3.3.2 Random Forest

RF is based on the theory of decision trees [43] usually used for classification (e.g. Lee et al. 2014) but that has also been demonstrated to be suitable for regression problems (e.g. [44,45]). A forest is a combination of tree predictors such that each tree depends on a vector of independently and randomly sampled values, or features, with the same distribution for all trees in the forest [36]. The error for the forest tends to converge as the number of trees

becomes large and depends on the strength of the individual trees and the correlation between them. Because of the random processes behind it, this method has been demonstrated to be robust with respect to outliers [36]. In the past, Herrera el al. [46] used RF to forecast hourly urban water demand in a city in south-eastern Spain. Chen et al. [47] used RF to forecast droughts and demonstrated that in this application RF outperformed other regression techniques. Recent studies also demonstrated the use of RF for making predictions of the fuel consumption of road vehicles based on on-board data [26]. Many software implemented the method and libraries like the `randomForest` R package [31] allow the user to apply RF by defining only a few parameters such as the number of trees in the forest (ntree) and the number of variables to consider and sample into each tree (mtry). The fact that each tree in the forest makes decisions based on a different subset of variables enables the forest to compare the error of trees containing only certain variables to the error of the complete model with all variables intact. This way the forest is able to rank each of the considered variables because of their importance in estimating the quantity of interest [36]. A higher number of trees usually implies higher precision and higher stability of the results, but at a higher computational cost. A rule of thumb is to set this value to the square-root of the total number of variables considered. Optimization of these parameters can improve the overall performance of the algorithm both in terms of reliability and calculation time.

3.3.3 Artificial Neural Networks

The third considered approach uses Artificial Neural Networks (ANN) [48,49]. ANN is a machine learning algorithm inspired by how the human brain processes information and is mostly used to estimate or approximate complex functions including nonlinear relationships that depend on a large number of variables [50]. Thanks to the possibility of parallel processing and the ability of ANN for adaptive learning, self-organization, and fault tolerance [50,34] the algorithm has been demonstrated to be a very powerful tool. Examples include the

use of ANN to predict medical outcomes (e.g. [51,52]), financial analysis for modelling stock performance (e.g. [53]), structural analysis (e.g. [54]) and transportation [55]. In the past, this technique has been applied for the estimation of the fuel consumption of aircraft [23,24], and more recently, of road vehicles [19,26] and to predict specific fuel consumption of diesel engines [56]. The main advantages of ANN are that it requires less formal statistical training than other machine learning methods and that it is able to implicitly detect complex nonlinear relationships between explanatory variables and the response [52]. There are many types of ANN, which use different types of neurons and activation functions. For this study, the adopted algorithm is the resilient propagation algorithm without backtracking (rprop-) [57] with logistic activation function. This has been chosen because it reduced the required calculation time and it requires fewer parameters to tune, compared to others. In the study, the rprop- neural network implemented in the `neuralnet` R package (Fritsch et al. 2016) was used.

3.4 Cross-Validation

In order to define more reliable models, which make predictions completely independent from how the available data are subset into training test datasets, ten-fold cross validation has been used in this study. The randomized splitting process has been repeated ten times and ten different models have been generated for each of the methods (SVM, RF, and ANN). The average performance (e.g. measurements of the accuracy of the generated models) is used for comparison purposes. Obtaining similar performance for each split of the data indicates that the available information is not affected by how the data is split. On the other hand, variations between data splits indicate lack of reliability in the model.

Some 96 ± 2.5% of data are used to apply ten-fold cross validation and generate the ten models. In particular, 75% of all data is used for training the model and 21% for validation. The remaining 4 ± 2.5% is then used in a second phase of the data analysis to test the

performance of the generated models. The two sets of data are randomly split with one condition; the cross validation set (96%) must not contain data from trucks in the testing set (4%). This way the testing set results contain completely new cases for the model and this checks the ability of the model to cope with completely new situations.

3.5 Test

Following the cross validation phase, the generated models need to be tested by analysing new data. Therefore, fuel consumption is computed for the 4% of data taken from the first phase of the analysis. This tests the reliability of the generated models and demonstrates the independence of the final estimates from the training data set.

3.6 Comparison of performance

Root mean square error (RMSE) and mean absolute error (MAE) are used to compare the performance and measure the accuracy of the generated models.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n}} \tag{3}$$

$$MAE = \left|\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n}\right| \tag{4}$$

where $\hat{y}_i$ is the i-th measurement of the dependant variable, y_i is the i-th prediction of the dependant variable, and n is the number of available records. Because the ten-fold cross validation technique generates ten models for SVM, RF, and ANN, the average RMSE and MAE of the testing phase are used to make a comparison with the performance of the linear regression model. Also, due to the fact that SVM, RF, and ANN require training before being able to make predictions, the computational time required for the ten-fold cross validation and test is compared between the machine learning methods.

3.7 Parametric analysis

Finally, a parametric analysis has been performed to understand how each of the considered variables impacts the fuel consumption of the considered fleet of trucks. The analysis consists of using the generated models to predict the fuel consumption of trucks for 50 different values of each of the variables considered in the model. This shows how each of the developed models approximates the relationships between fuel consumption and the explanatory variables and tests the sensitivity of each of them. The 50 values were chosen to be evenly distributed between the 5^{th} and 95^{th} percentile values of the distribution of the considered variable. While the value of one variable changes all others are set to their average.

4. Results

From the analysis of the Pearson's correlation coefficients, it is possible to conclude that ten out of 44 available measurements are correlated to fuel consumption, having a correlation coefficient higher than 0.10 and being independent of each other. These are the gross vehicle weight (*Gross.Weight*), the road gradient (*Geom.Gradient*), the vehicle speed (*Speed.AVG*), the torque % at the start of the travel (*Torque.Start*), the torque % at the end of the travel (*Torque.End*), the revolutions (*Revs.Start*) at the start of the travel, the average acceleration (*Acceleration*), the selected gear (*Gear*), the cruise control (*Cruise.Control*) (on/off, 1/0), and the radius of curvature of the road (Geom.Abs_*Radius*). Figure 1 shows their distributions for the considered case study. From the figure, it is possible to see that generally, the distributions of the continuous variables have shapes typical of normal distributions and can, therefore, be used in the linear regression analysis. However, there are exceptions. For example, the radius of curvature (*Geom.Abs_Radius*), has a left skewed distribution, and this is because the recording software assigns all road sections with radius of curvature above

2000m a value equal to 2000m (because they are considered nearly straight and the value therefore, is considered of no interest to the asset manager). It can be seen that the majority of road sections are nearly straight, which is typical of motorways.

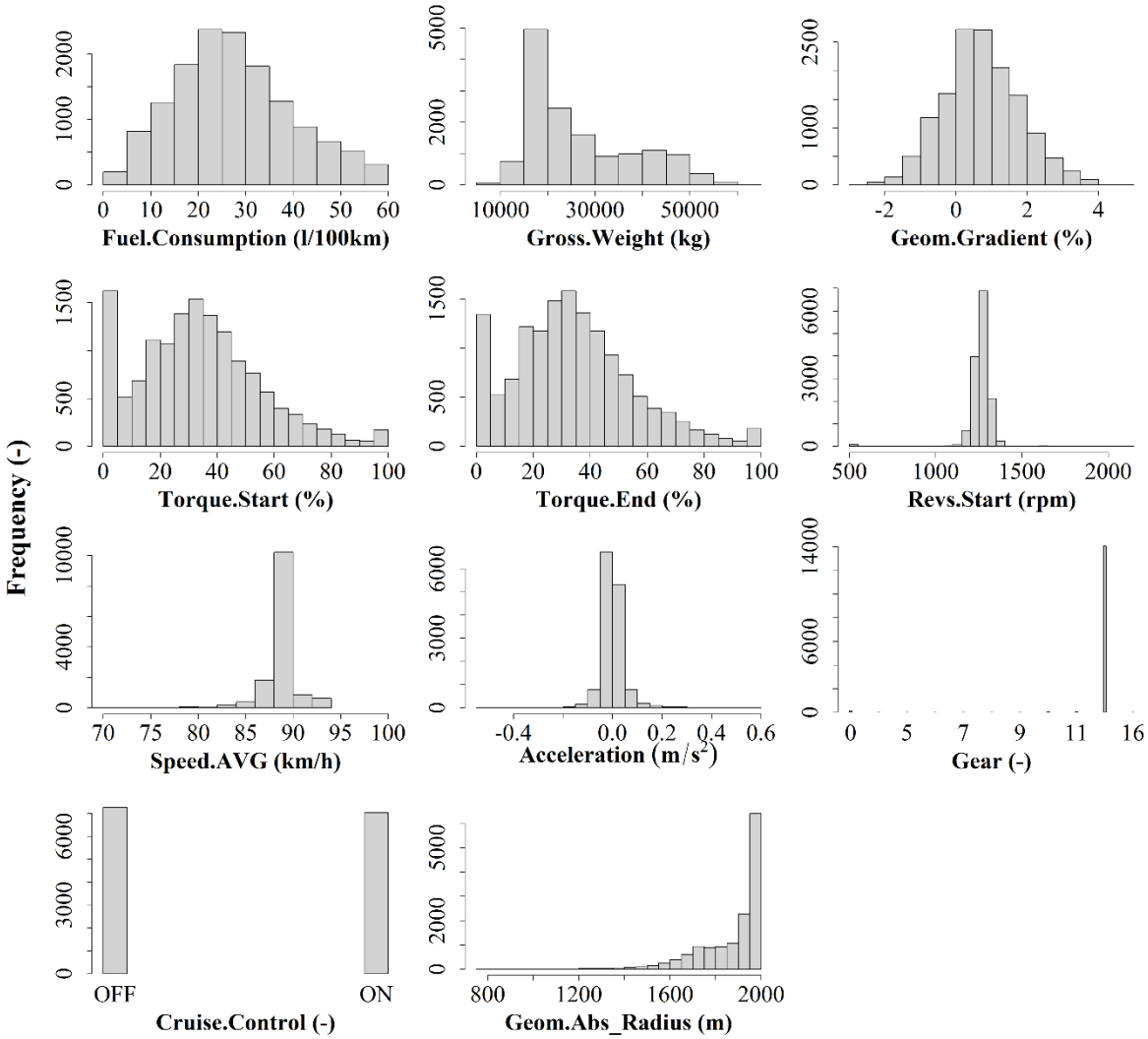

Fig. 1. *Distribution of the variables that show high correlation (Pearson's coefficient > 0.10) with Fuel.Consumption.*

The measurements of torque (both *Torque.Start* and *Torque.End*) have a high number of values towards zero. These data are possibly due to faulty sensors, however this cannot be confirmed and, for this reason, they cannot be excluded from the analysis as outliers. Furthermore, due to the fact that only records at a near constant speed are considered, it is possible to see that only a limited range of vehicle speeds (*Speed.AVG*) is considered, from around 70 to 95 km/h, which is typical for motorways. Consequently, only a few gears are

used (*Gear*), mainly gear 12, producing a narrow range of engine revolutions (*Revs.Start*) (from 1000 to 1300 rpm). For the same reasons, a low level of acceleration (*Acceleration*) results. P-values are computed for the remaining 10 variables to test whether or not they are significant for predicting the fuel consumption of the fleet of trucks considered, using a linear regression model (Table 2). From the analysis of the computed p-values, only eight out of the ten variables show a significant correlation with fuel consumption (p-value < 0.05). Therefore, it results that the activation of cruise control (*Cruise.Control*) should be excluded from the data analysis.

Table 2 shows the p-values of correlated variables.

Name of the variable	**Computed p-value**	**Significance**
Gross.Weight	$< 2 \cdot 10^{-16}$	> 99%
Geom.Gradient	$< 2 \cdot 10^{-16}$	> 99%
Speed.AVG	$< 2 \cdot 10^{-16}$	> 99%
Torque.Start	$< 2 \cdot 10^{-16}$	> 99%
Torque.End	$< 2 \cdot 10^{-16}$	> 99%
Revs.Start	$< 2 \cdot 10^{-16}$	> 99%
Acceleration	$< 2 \cdot 10^{-16}$	> 99%
Gear	$< 2 \cdot 10^{-16}$	> 99%
Cruise.Control	0.121	0%
Geom.Abs_Radius	0.165	0%

Also, the radius of curvature of the road (*Geom.Abs_Radius*) should be excluded due to the fact that its p-value is higher than 0.05 and therefore considered as of low significance for the case study. This is reasonable since this study considers generally straight roads and the result may also be influenced by the method of recording. However, looking at Figure 2, which shows the results of the analysis of the adjusted-R^2, it is possible to see that the inclusion of both the *Cruise.Control* and *Geom.Abs_Radius* actually helps in increasing the adjusted-R^2 meaning that this variable can help in improving the accuracy of the developed regression models. The graph shows that all ten (of the initial 44) variables can be considered as playing an important role in estimating fuel consumption.

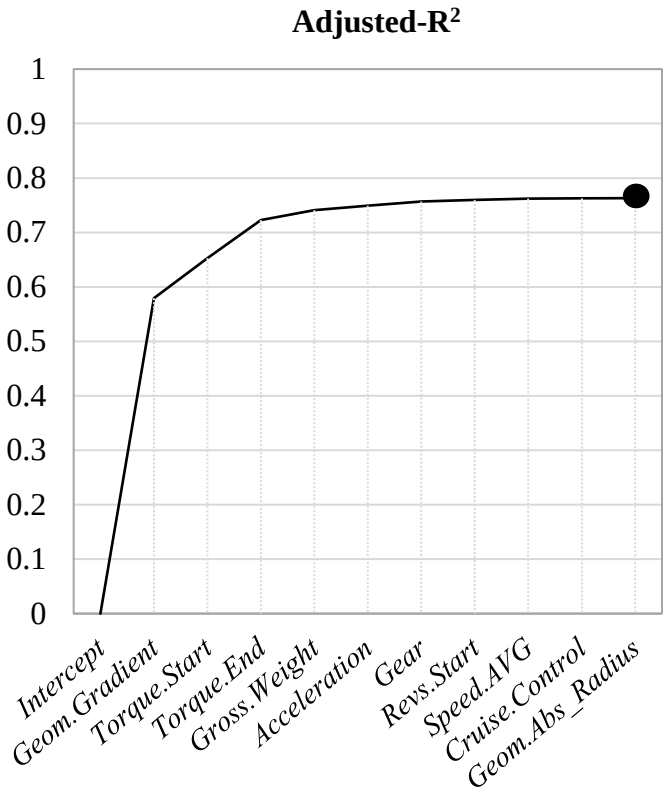

Fig. 2. *Increase in the adjusted-R^2 by including new variables in the model.*

A Lasso regression was performed on the variables to confirm that the correlation between them and fuel consumption is significant for the considered case study.

Table 3 shows the Lasso coefficients of the selected variables.

Variable	**Lasso coefficient**
Gross.Weight	0.143
Geom.Gradient	0.524
Speed.AVG	-0.034
Torque.Start	0.248
Torque.End	0.275
Revs.Start	0.042
Acceleration	-0.143
Gear	-0.049
Cruise.Control	0.0049
Geom.Abs_Radius	0.0057

From the analysis of the Lasso coefficients (Table 3), it is possible to see that the algorithm is not able to reduce to zero any of the parameters and all the selected variables can be considered significant including *Cruise.Control* and *Geom.Abs_Radius*.

Finally, the Random Forest algorithm was used to rank the variables for a final test. As an 'increase in node purity' makes the developed models able to make more accurate estimates [36], it is possible to see (Figure 3) that *Cruise.Control* and *Geom.Abs_Radius* are confirmed to play an important role in the prediction of fuel consumption as more important than other variables (i.e. *Speed.AVG*, *Revs.Start*, and *Gear*) that the p-values and adjusted-R^2 identified.

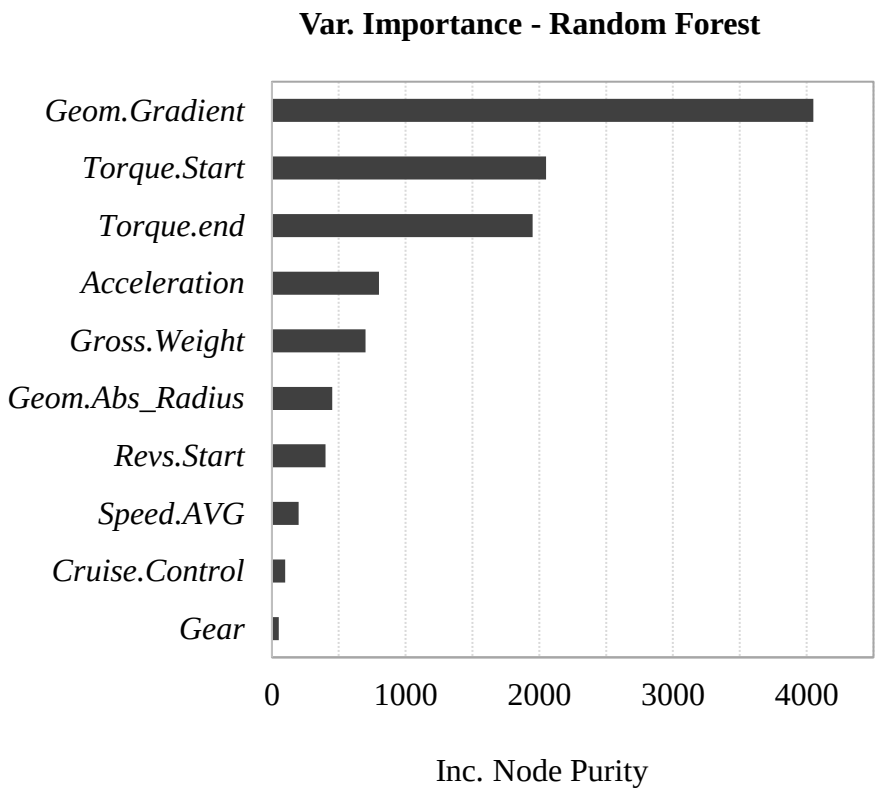

Fig. 3. *Variable ranked by the Random Forest algorithm by their importance.*

However, the activation of cruise control (*Cruise.Control*) is measured as a Boolean variable (0/1) and that does not allow a linear regression to handle it properly. This is probably why the associated p-value results being poorly significant while other statistics disagree. Also, the linear regression technique is not suitable to analyse the measurement of the radius of

curvature of the road since the *Geom.Abs_Radius* represents all radii over 2,000m as having this value. This makes the *Geom.Abs_Radius* a non-continuous variable and this is why the p-value associated to this measurement result to be non-significant. Although the *Gear* is also a non-continuous variable its associated p-value results to be significant and all the used statistics agree on this. Because the produced evidence does not completely agree on the fact that *Cruise.Control* and *Geom.Abs_Radius* should be included in the linear regression model, in order to avoid overfitting and make the generated models completely comparable, a decision was made to exclude these two variables from the following regression analysis.

Therefore, eight variables are included in the developed models. These are: the *Gross.Weight*, the *Geom.Gradient*, the *Speed.AVG*, the *Torque.Start*, the *Torque.End*, the *Revs.Start*, the *Acceleration*, and the *Gear*.

From a multiple linear regression of the selected variables the generated model takes the form:

$$\begin{aligned} FC \;=\; & 20.3 + 1.6 \times 10^{-4} \times Gross.Weight + 6.04 \times Geom.Gradient - 0.15 \\ & \times Speed.AVG + 0.14 \times\; Torque.Start + 0.16 \times Torque.End + 6.8 \\ & \times\; 10^{-3} \times Revs.Start +\; 34.04 \times Acceleration + 0.56\; \times Gear \end{aligned} \tag{5}$$

Regarding the SVM, it is difficult to make a pictographic representation of the model due to the fact that multiple hyperplanes with very complex expressions are generated by the algorithm. For this reason, no equation or graph representing the SVM model is reported.

For the RF, the generated model takes the form of multiple decision trees (that comprise the forest). Example of part of one of the 1,000 trees in the generated models is given in Figure 4.

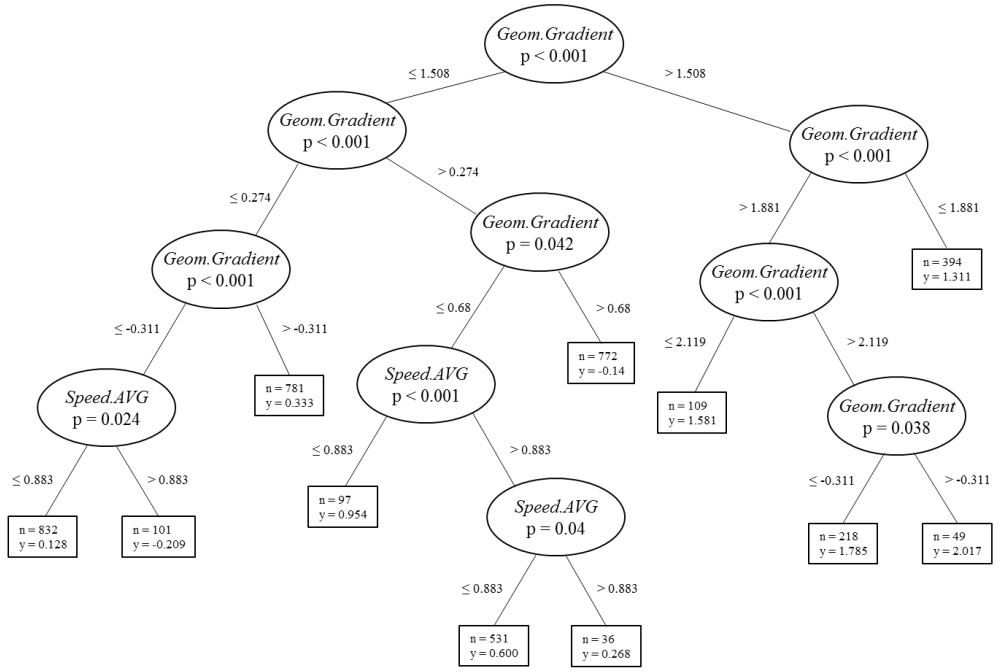

Fig. 4. *Structure of part of one of the 1,000 trees used in the Random Forest model.*

The number of trees and the number of variables to be used in each of the trees have been selected as the values which optimize RMSE, MAE, and computational time required.

For the ANN the structure of the generated model takes the form in Figure 5.

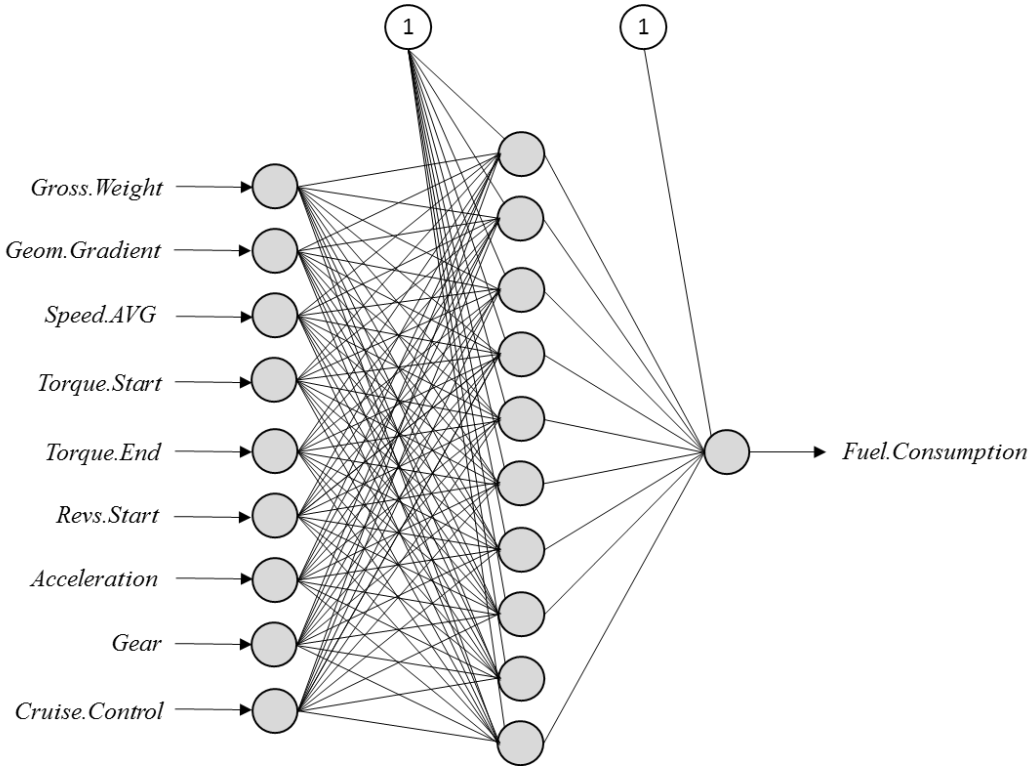

Fig. 5. *Structure of the generated Artificial Neural Network model.*

The structure of the ANN model was decided by evaluating the performance of the combination of one and two hidden layers with maximum 12 nodes each. Also, in this case, the structure showing the lowest RMSE and MAE, and reasonable computational time required (< 1 min per training) has been selected. This resulted in using an ANN with a single hidden layer and ten nodes. The R^2, RMSE, and MAE statistics were calculated for the cross-validation and testing phase of the regression analyses and compared. The R^2 for SVM, RF, and ANN is calculated for the testing phase of the regression analysis. RMSE and MAE are instead calculated in both the cross-validation and testing phase. This is to test the robustness of the generated models. In fact, the constant performances of the model in the cross-validation and testing phase of the regression analysis, highlights the quality of the models. The calculation times (including the time needed for cross-validation of the SVM, RF, and ANN models) are also shown.

Table 4 summarizes the performance of the generated models.

Model	**R^2**	**$RMSE_{test}$**	**MAE_{test}**	**$RMSE_{cv}$**	**MAE_{cv}**	**cv Time**
Linear regression	0.763	-	-	6.02	4.42	-
Support Vector Machine (SVM)	0.821	5.30	3.69	5.20	3.53	~ 5 mins
Random Forest (RF)	0.835	5.12	3.58	4.86	3.38	~ 26 mins
Artificial Neural Network (ANN)	0.814	5.43	3.91	5.18	3.50	~ 9 mins

Figure 6 shows how the models are able to predict real measurements. In particular, four plots in the figure shows the fit of measured versus predicted values of fuel consumption for each of the developed models. A low grade of transparency is given to the datapoints in the plot of the fit made by the linear regression in order to make possible identifying higher density areas. Three different colours are given in the plots of the fits given by the machine learning models: light grey is for the training dataset, grey is for the test set, and dark grey for the validation set. For each of the machine learning methods applied only results of one of the ten crossvalidated models is shown. The figure shows how machine learning methods are

able to reduce variance thanks to their resilience to outliers. From an analysis of Figure 6 and Table 4, it is possible to conclude that all machine learning models developed in the study perform better than the linear regression. The RF model provides the best predictions of fuel consumption showing the highest R^2, the lowest RMSE and MAE both in the cross-validation and testing phase of the regression analysis, and the lowest variance among all the generated models.

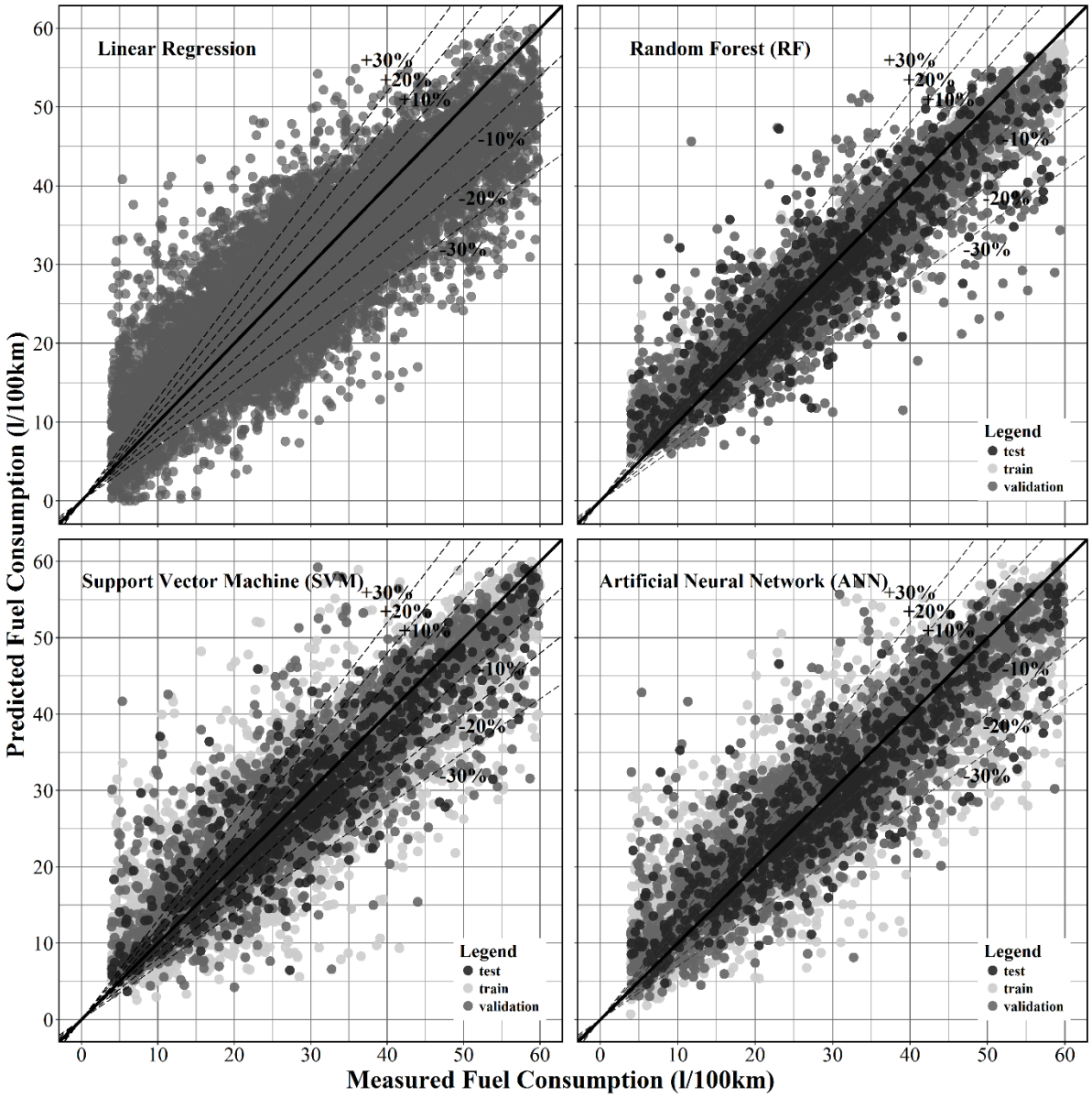

Fig. 6. *Comparison of the predicted and measured data for the generated models. From the top left and going clockwise: linear regression, support vector machine (SVM), artificial neural network (ANN), and random forest (RF) regressions.*

However, RF is also the technique that requires the longest calculation time for training the model and this may be a disadvantage of the algorithm. Furthermore, it is possible to see from Figure 6 that the RF is not so good in predicting extreme values of fuel consumption, with both high and particularly lower values tending to the higher confidence limits, while the SVM performs better in this respect. Finally, a parametric analysis has been performed to see how machine learning addresses the impacts of each of the considered variables on fuel consumption and Figure 7 shows the results. Similar trends are shown by the SVM, RF, and ANN models. However, in order to make the plot clearer, and for brevity, only the results of the linear regression and ANN regression model are presented in this paper. The figure shows that the trends plotted by the linear regression and the ANN are different in some respects. Unreliable predictions of fuel consumption are made by the linear model for very low values of road gradient. Extrapolating the linear model could lead to the conclusion that travelling on large downhill gradients can result in negative fuel consumption. This is not possible and the asymptotic approximation made by the machine learning algorithms more realistic. It is interesting to note that the ANN is able to detect an optimal gear (gear 12) for driving at a fairly constant speed (around 90 km/h) on a motorway. Another interesting point is that lower fuel consumption is associated with higher speed and acceleration. Although, at first, these findings would sound strange, it must be considered that the range of speeds and accelerations considered in the study is very narrow and that the effect of road gradient may be prevailing. In particular a negative gradient (downhill) could lead to higher speeds and positive accelerations yet lower fuel consumption, while positive gradients (uphill) would cause instead lower speeds, negative accelerations and higher fuel consumption. This might also explain the discontinuity detected by the machine learning algorithms between negative and positive accelerations, as the model tries to explain two separate phenomena; however, this requires further investigation.

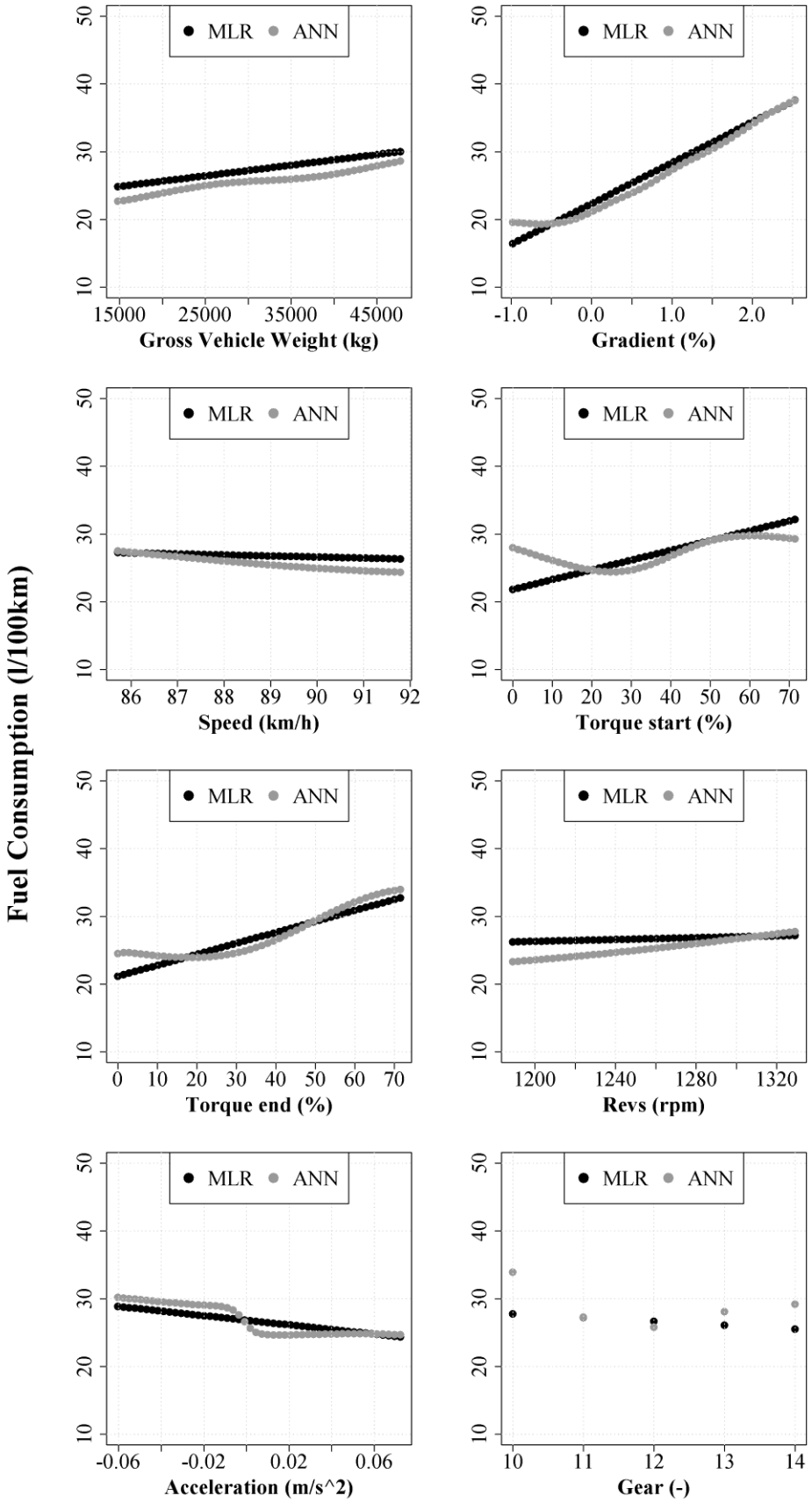

Fig. 7. *Parametric analysis for the multiple linear regression and ANN generated models.*

5. Conclusions and Future Work

This paper investigates the estimation of fuel consumption of a large fleet of trucks based on truck sensor telematic data and road geometry. The performances of three machine learning regression techniques are compared to a multiple linear regression of the explanatory variables. Among the 44 measurements initially available, only eight were included in the generated models based on the Pearson's correlation coefficients, the p-values, the adjusted-R^2, the Lasso regression and the Random Forest algorithm. These are the gross vehicle weight, the road gradient, the vehicle speed, the initial and final values of torque percentage, the initial revolutions of the engine, the acceleration and the selected gear. Although the activation of the cruise control and the radius of curvature of the road may be able to improve the performance of the machine learning models (as they have been identified to be significant by the Lasso and the Random Forest algorithm), these are not considered in the study bacause they cannot be classified as continuous variables and, from the analysis of the associated p-values, they have been identified to be poorly significant for the considered case study. Excluding them from the regression analysis allows the generated models to include the same variables, avoids overfitting of the linear regression and permits a full comparison of the performance of the developed models.

The study shows how, under controlled conditions (e.g. approximately constant speed, only one vehicle type considered, etc.), SVM, RF, and ANN perform better than a simple multiple linear regression in predicting the fuel consumption of the considered fleet of trucks. Therefore, it can be concluded that although the present study only focused on a simplified case, the effect of nonlinearity is significant and cannot be considered negligible when modelling the fuel consumption of trucks. The R^2 values imply that further work is needed to include more variables in the developed models and it will be interesting to extend this work to include the effect of the activation of cruise control, the radius of curvature of the road

(excluded in this study) and analysing a wider range of conditions for different vehicle types. When factors such as higher accelerations, braking, wind speed, etc. are introduced into the models, further nonlinearity is likely to result, and the use of machine learning techniques in estimating fuel consumption will become more important.

Results of this study showed that, among the considered techniques, RF gives the best values of RMSE and MAE for the cross-validation and testing sets. Another possible benefit of using RF is the possibility of computing variable importance [36]. In future work, this technique could be used instead of p-values to select the explanatory variables to include in the regression analysis while accounting for any complex nonlinear relationship. However, SVM and ANN also demonstrated a good level of accuracy in making estimations. Another benefit of using SVM or ANN instead of the RF algorithm is that the cross-validation process of SVM and ANN was about three times faster than for the RF. For these reasons, it is not possible to conclude which of the investigated machine learning methods performs better overall. It is possible to say that machine learning methods are the better alternative compared to linear regression models to estimate fuel consumption. They detect complex nonlinear relationships that exist even in relatively simple cases like that analysed in this paper.

The approach described here can be used by manufacturers to estimate the actual GHG emissions produced by their vehicles in real driving conditions and by road agencies and designers to estimate GHG emissions resulting from the use phase of the road infrastructure due to road geometry. Once the fuel consumption estimates are known it is possible to use existing tools (e.g. [58]) to estimate the consequent GHG emissions per liter of fuel used or the equivalent amount of CO_2 released into the atmosphere. For this reason, the methodology and models introduced in this paper may have an impact on the decision-making of vehicle manufacturers, standards committees, road designers and asset managers. However, although

initial results seem to be promising more work is still to be done. Validation of these results for a wider range of vehicles, including more variables, e.g. the effect of the air temperature, wind speed, or driver behavior [59,60], etc., will improve the applicability of the study.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 642453. The authors would like to thank Ian Dickinson, from Microlise Ltd, for allowing the use of anonymized data from their databases and helping in understanding and analyzing the data, and Highways England for permitting access to HAPMS. The authors also want to thank Helen Viner, Emma Benbow, David Peeling and Thomas Buckland, from TRL Ltd, for their help and support in the process of extracting the road data and in interpreting some of the results of the study.

References

[1] Department of Business Energy and Climate Change, 2015 UK Greenhouse Gas Emissions , Final Figures Statistical Release, 2017.

[2] J. Lin, D.A. Niemeier, An Exploratory Analysis Comparing a Stochastic Driving Cycle to California's Regulatory Cycle, Atmos. Environ. 36 (2002) 5759–5770.

[3] S.H. Kamble, T. V. Mathew, G.K. Sharma, Development of real-world driving cycle: Case study of Pune, India, Transp. Res. Part D Transp. Environ. 14 (2009) 132–140.

[4] S. Shahidinejad, E. Bibeau, S. Filizadeh, Statistical development of a duty cycle for plug-in vehicles in a North American urban setting using fleet information, IEEE Trans. Veh. Technol. 59 (2010) 3710:3719.

[5] H. Tong, H. Tung, W. Hung, N. H., Development of driving cycles for motorcycles and light-duty vehicles in Vietnam. Atmospheric Environment, Atmos. Environ. 45

(2011) 5191–5199.

[6] A. Marotta, M. Tutuianu, Europe-centric light duty test cycle and differences with respect to the WLTP cycle, JRC Sci. Policy Reports. Tech. Rep. - Task 7. Eur. Comm. Inst. Energy Transp. (2012).

[7] B. Ciuffo, A. Marotta, M. Tutuianu, K. Anagnostopoulos, G. Fontaras, J. Pavlovic, S. Serra, S. Tsiakmakis, The development of the World-wide Harmonized Test Procedure for Light Duty Vehicles (WLTP) and the pathway for its implementation into the EU legislation, Transp. Res. Board Annu. Meet. 15–4935 (2015). doi:http://dx.doi.org/10.13140/RG.2.1.3175.8562.

[8] R. Giannelli, E. Nam, K. Helmer, T. Younglove, G. Scora, M. Barth, Heavy-duty diesel vehicle fuel consumption modeling based on road load and power train parameters, SAE Int. (2005) 13. doi:http://dx.doi.org/10.4271/2005-01-3549.

[9] N. Clark, G. Thompson, O. Delgado, Modeling Heavy-duty Vehicle Fuel Economy Based on Cycle Properties, West Virginia University, 2009.

[10] O.F. Delgado, N.N. Clark, G.J. Thompson, Modeling Transit Bus Fuel Consumption on the Basis of Cycle Properties, J. Air Waste Manage. Assoc. 61 (2011) 443–452. doi:http://dx.doi.org/10.3155/1047-3289.61.4.443.

[11] K. Chatti, I. Zaabar, Estimating the Effects of Pavement Condition on Vehicle Operating Costs, NCHRP Rep. 720. (2012).

[12] G. Bifulco, F. Galante, L. Pariota, M. Spena, A Linear Model for the Estimation of Fuel Consumption and the Impact Evaluation of Advanced Driving Assistance Systems, Sustainability. 7 (2015) 14326–14343. doi:10.3390/su71014326.

[13] N.P.D. Martin, J.D.K. Bishop, R. Choudhary, A.M. Boies, Forecasting Passenger Fleet Fuel Consumption – A New Methodology to Include Uncertainty Analysis, Transp. Res. Board, TRB 94th Annu. Meet. Compend. Pap. Washington, DC. (2015).

[14] E. Beuving, T. De Jonghe, D. Goos, T. Lindahl, A. Stawiarski, Environmental Impacts and Fuel Efficiency of Road Pavements, (2004).

[15] L. Guzzella, A. Sciaretta, Vehicle Propulsion Systems, Introduction to Modeling and Optimization, Springer Science+Business Media, Zurich, 2015. doi:http://dx.doi.org/10.1007/978-3-642-35913-2.

[16] H.R. Kerali, J.B. Odoki, E.E. Stannard, Overview of HDM-4, Volume 1, Highw. Dev. Manag. Ser. Int. Study Highw. Dev. Manag. (2006).

[17] E.K. Nam, R. Giannelli, Fuel Consumption Modeling of Conventional and Advanced Technology Vehicles in the Physical Emission Rate Estimator (PERE) - Draft, Tech. Rep. EPA. (2005) 124.

[18] M. Goo Lee, Y. Kuk Park, K. Kwon Jung, J. Jae Yoo, Estimation of Fuel Consumption using In-Vehicle Parameters, Int. J. U- E- Serv. Sci. Technol. 4 (2011) 37–46. doi:10.6109/jkiice.2011.15.12.2582.

[19] 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.

[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Adv. Neural Inf. Process. Syst. (2012) 1–9. doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007.

[21] D.A. Clifton, K.E. Niehaus, P. Charlton, G.W. Colopy, Health Informatics via Machine Learning for the Clinical Management of Patients, Yearb. Med. Inform. 10 (2015) 38–43. doi:http://dx.doi.org/10.15265/IY-2015-014.

[22] S. Maes, K. Tuyls, B. Vanschoenwinkel, Credit Card Fraud Detection Using Bayesian and Neural Networks, Maciunas RJ, Ed. Interact. Image-Guided Neurosurgery. Am. Assoc. Neurol. Surg. (1993) 261–270. doi:http://dx.doi.org/10.1.1.18.5377.

[23] G.D. Schilling, Modeling Aircraft Fuel Consumption with a Neural Network, Comput.

Sci. (1997).

[24] A.A. Trani, G. Schilling, H. Baik, A. Seshadri, A Neural Network Model to Estimate Aircraft Fuel Consumption, in: AIAA 4th Aviat. Technol. Integr. Oper. Forum, American Institute of Aeronautics and Astronautics, Chicago, Illinois, 2004: p. 24.

[25] X. Xu, Y. Zhao, Prediction of fuel consumption per 100km for automobile engine based on Gaussian processes machine learning, in: Int. Conf. Mech. Eng. Green Manuf. 2010, MEGM 2010, 2010: pp. 1951–1955. doi:10.4028/www.scientific.net/AMM.34-35.1951.

[26] H. Almér, Machine learning and statistical analysis in fuel consumption prediction for heavy vehicles, KTH, Sweden, 2015.

[27] SAE International, SAE J1939-71, Vehicle Application Layer - Surface Vehicle Recommended Practice, SAE Int. Stand. (2016). http://standards.sae.org/j1939/71_201610/.

[28] J. Friedman, T. Hastie, N. Simon, J. Qian, R. Tibshirani, R Package: glmnet, Version 2.0-13, (2017).

[29] M. Kuhn, R Package: caret - Classification and Regression Training, Version 6.0-76, 2017. https://github.com/topepo/caret/.

[30] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, C.-C. Lin, R Package: e1071, Version 1.6-8, 2017. https://cran.r-project.org/web/packages/e1071/e1071.pdf.

[31] A. Liaw, M. Wiener, L. Breimann, A. Cutler, R Package: randomForest, Version 4.6-12, (2017). doi:http://dx.doi.org/10.5244/C.22.54.

[32] S. Fritsch, F. Guenther, M. Suling, S.M. Mueller, R Package: neuralnet, Version 1.33, (2016).

[33] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, J. R. Stat. Soc. Ser.

B. 58 (1996) 267–288. doi:http://dx.doi.org/10.1111/j.1467-9868.2011.00771.x.

[34] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Springer Science+Business Media New York, New York, NY, 2013. doi:10.1007/978-1-4614-7138-7.

[35] Y. Zhang, W. Guo, C. Edu, On the Consistency of Feature Selection With Lasso for Non-linear Targets, in: Proc. 33rd Int. Conf. Mach. Learn., JMLR: W&CP, New York, NY, USA, 2016.

[36] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32. doi:http://dx.doi.org/10.1023/A:1010933404324.

[37] C. Cortes, V. Vapnik, Support-Vector Networks, Mach. Learn. 20 (1995) 273–297. doi:10.1023/A:1022627411411.

[38] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Networks. 10 (1999) 988–999. doi:10.1109/72.788640.

[39] S.R. Gunn, Support Vector Machines for Classification and Regression, Southampton, 1998. doi:10.1039/B918972F.

[40] L.J. Cao, F.E.H. Tay, Support vector machine with adaptive parameters in financial time series forecasting, IEEE Trans. Neural Networks. 14 (2003) 1506–1518. doi:http://dx.doi.org/10.1109/TNN.2003.820556.

[41] M.S. Khan, P. Coulibaly, Application of Support Vector Machine in Lake Water Level Prediction, J. Hydrol. Eng. 11 (2006) 199–205. doi:http://dx.doi.org/10.1061/(ASCE)1084-0699(2006)11:3(199).

[42] C.-C. Chang, C.-J. Lin, libsvm, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27. doi:http://dx.doi.org/10.1145/1961189.1961199.

[43] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, 1984. doi:http://dx.doi.org/10.1371/journal.pone.0015807.

[44] A. Deloncle, R. Berk, F. D'Andrea, M. Ghil, Weather Regime Prediction Using Statistical Learning, J. Atmos. Sci. 64 (2007) 1619–1635. doi:http://dx.doi.org/10.1175/JAS3918.1.

[45] L. Khaidem, S. Saha, S.R. Dey, Predicting the direction of stock market prices using random fores, CoRR. (2016). http://arxiv.org/abs/1605.00003.

[46] M. Herrera, L. Torgo, J. Izquierdo, R. Pérez-García, Predictive models for forecasting hourly urban water demand, J. Hydrol. 387 (2010) 141–150. doi:http://dx.doi.org/10.1016/j.jhydrol.2010.04.005.

[47] J. Chen, M. Li, W. Wang, Statistical uncertainty estimation using random forests and its application to drought forecast, Math. Probl. Eng. 2012 (2012) 1–13. doi:http://dx.doi.org/10.1155/2012/915053.

[48] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (1943) 115–133. doi:http://dx.doi.org/10.1007/BF02478259.

[49] P. Werbos, Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences, Harvard University, Washington (DC), 1974. http://www.citeulike.org/group/1938/article/1055600.

[50] C. Stergiou, D. Siganos, Neural Networks, (1995). https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#What is a Neural Network (accessed October 4, 2016).

[51] H.B. Burke, Artificial neural networks for cancer research: Outcome prediction, Semin. Surg. Oncol. 10 (1994) 73–79. doi:10.1002/ssu.2980100111.

[52] J. V. Tu, Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, J. Clin. Epidemiol. 49 (1996) 1225–1231. doi:http://dx.doi.org/10.1016/S0895-4356(96)00002-9.

[53] A.N. Refenes, A. Zapranis, G. Francis, Stock performance modeling using neural networks: a comparative study with regression models, Neural Networks. 7 (1994) 375–388. doi:10.1016/0893-6080(94)90030-2.

[54] A.A. Chojaczyk, A.P. Teixeira, L.C. Neves, J.B. Cardoso, C. Guedes Soares, Review and application of Artificial Neural Networks models in reliability analysis of steel structures, Struct. Saf. 52 (2015) 78–89. doi:http://dx.doi.org/10.1016/j.strusafe.2014.09.002.

[55] M. Dougherty, A review of neural networks applied to transport, Transp. Res. - C. 3 (1995) 247–260.

[56] A. Parlak, Y. Islamoglu, H. Yasar, A. Egrisogut, Application of artificial neural network to predict specific fuel consumption and exhaust temperature for a Diesel engine, Appl. Therm. Eng. 26 (2006) 824–828. doi:10.1016/j.applthermaleng.2005.10.006.

[57] M. Riedmiller, H. Braun, A direct adaptive method for faster backpropagation learning: The RPROP algorithm, in: IEEE Int. Conf. Neural Networks - Conf. Proc., 1993: pp. 586–591. doi:10.1109/ICNN.1993.298623.

[58] IPCC, 2006 IPCC Guidelines for National Greenhouse Gas Inventories, Prepared by the National Greenhouse Gas Inventories Programme, IGES, Japan, 2006.

[59] G.P. Figueredo, P.R. Quinlan, M. Mesgarpour, J.M. Garibaldi, R.I. John, A Data Analysis Framework to Rank HGV Drivers, IEEE Conf. Intell. Transp. Syst. ITSC. October, 2 (2015) 2001–2006. doi:http://dx.doi.org/10.1109/ITSC.2015.324.

[60] G.P. Figueredo, U. Agrawal, J.M. Mase, M. Mesgarpour, C. Wagner, D. Soria, J.M. Garibaldi, P. Siebers, R.I. John, Identifying Heavy Goods Vehicle Driving Styles in the United Kingdom, IEEE Trans. Intell. Transp. Syst. (2017) (accepted).