Diesel Engine Particulate Emissions Prediction Using a Machine Learning Algorithm

Danilo Gomes Dellaroza[1,3]
Marco Antônio Luersen[1]
Germano Custódio Lodi[2]
[1] Universidade Tecnológica Federal do Paraná
[2] Universidade Positivo
[3] Volvo do Brasil

ABSTRACT

One of the requirements of diesel engines certification process is that the engine do not exceed the specific particulate emissions limit in cycle ($Spec\_Partcycle$). To calculate the $Spec\_Partcycle$, a manual process is required. The total amount of mass impregnated in the particulate filter is obtained by weighting the filter on a balance after each test. Therefore, it is not possible to obtain it without human intervention. In order to allow test rigs to operate in automatic mode, without an operator conducting the tests, an automatic way of calculating $Spec\_Partcycle$ is required. Thus, the aim of this work is to develop a $Spec\_PartCycle$ prediction model that does not require human intervention. For this, a machine learning approach, based on the random forests algorithm, is used. Data collected from 2012 to 2019 from three test cells of 11 and 13 liters diesel engines of an automotive company, summing up 2500 valid test results, are used as input. This data are employed to train the algorithm and build a prediction model. The prediction model is then validated using another 72 validation tests results. The accuracy of the final model considering a confidence interval of 95% is ±3,00 mg/kWh for the European Transient Cycle, and ±1,96 mg/kWh for the European Stationary Cycle.

INTRODUCTION

Vehicle manufactures must meet several requirements to get the permission to sell their products, for example, safety, environmental or weight requirements. One of the most regulated markets is the diesel vehicles. Today, diesel engines are preferred in several industries because of their higher fuel economy and better efficiency [1]. However, the exhaust emissions of these engines have been adversely affecting humanity and habitat for many years [2]. Due to this reason, restricted regulations are imposed to diesel engine vehicle manufactures.

During the engine development and certification process, a list of parameters is verified and must comply with legal requirements. In Brazil, the current environmental legislation that regulates emissions in heavy diesel vehicles is CONAMA Resolution n. 403 [3]. According to this regulation, emissions limit for heavy duty diesel engines must be in accordance with Table 1. Two types of tests must be carried out [3]: the European Stationary Cycle (ESC) and the European Transient Cycle (ETC). In ESC test, the engine is tested on a dynamometer over a sequence of steady-state modes [4]. In ETC test, three different driving conditions are represented: urban, rural and motorway driving [4].

Table 1 – Emission limits for heavy-duty diesel according to CONAMA Resolution n. 403 [3].

Test	CO	THC	NMHC	NOx	PM	Smoke
			g/kWh			m^{-1}
ESC	1.5	0.46	-	2.0	0.02	0.5
ETC	4.0	-	0.55	2.0	0.03	-

CO = carbon monoxide (g/kWh)
THC = total hydrocarbons (g/kWh)
NMHC = non methane hydrocarbons (g/kWh)
NOx = nitrogen oxides (g/kWh)
PM = particulate matter (g/kWh)
Smoke = Smoke Opacity (m^{-1})

In engine test rigs, usually as soon as the test cycle (ESC or ETC) is over, the results for CO, THC, NHMC NOx and Smoke are calculated automatically. All the automation to get these results is already developed and several suppliers provide equipment that contributes to this. The only parameter that requires human intervention is the particulate matter measurement (PM or $Spec\_Partcycle$). Due to this reason, usually an operator is required to run the tests, which is not a desired condition because it involves additional costs and the manual tasks after each test reduces the efficiency of the laboratory.

In order to have the particulate mass after a test cycle, a complex system is necessary. It is a particulate sampling system based on the partial flow method and variable dilution. It satisfies the specifications reported in all main regulations for the homologation of diesel engines from light up to heavy duty, both on-road and off-road. Figure 1 shows a scheme of how it works [5].

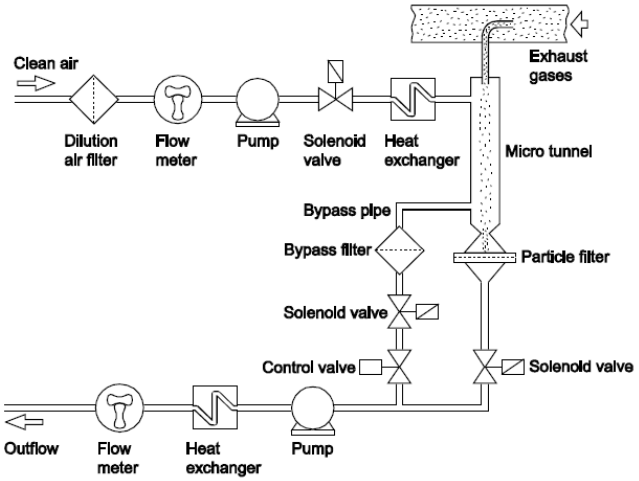

Figure 1- PSS-20 working scheme [5].

Before starting and after finishing the test, the particle filter must be weighted. Figure 2 shows how two particle filters at different positions look like after an ETC cycle. One of the represented filters is located at the engine out and the other after the Diesel Particulate Filter (DPF).

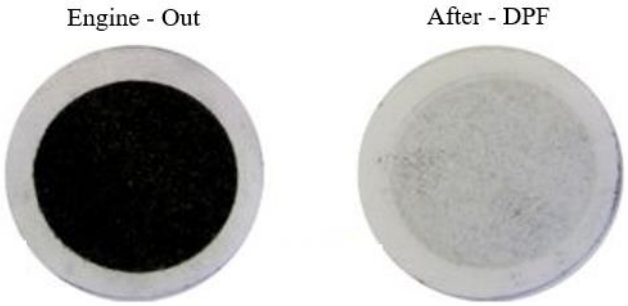

Figure 2 - Filters with PM samples [5].

By definition, "diesel particulates" are measured as any material deposited on the filter from the dilute exhaust gases sampled. It should be remembered that, because of this definition, diesel particulate matter includes not only solids but also liquid material which would condense in the form of mist or droplets at the above temperature, such as sulphuric acid or high-boiling hydrocarbons [6].

Figure 3 shows the composition of diesel particulate matter for a heavy-duty diesel engine submitted to a US FTP transient cycle, 500 ppm of maximum sulphur in fuel.

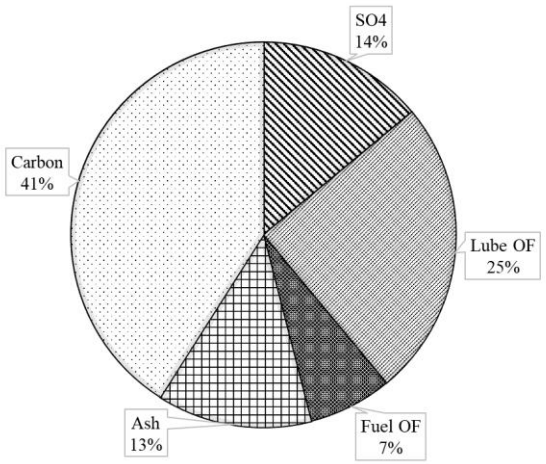

Figure 3 - Composition of diesel particulate matter [6].

The total mass impregnated on the filter of particulate measurement system must be weighted after a test cycle, which value is used to calculate the $Spec\_Partcycle$, according to:

$$SpecPartcycle = \frac{FlwPartcycle \times 3600}{Pwr\_CycleNet} \quad (1)$$

where $Pwr\_CycleNet$ (kWh) is the measured cycle net power and $Flw\_Partcycle$ (mg/s) is the cycle mass flow of particulates from microtunnel which is obtained by:

$$FlwPartcycle = \frac{MaPartfTot \times \sum Flw\_EqvCVS \times WF \times 1000}{\sum MaPartFilter} \quad (2)$$

where Ma_Partf_Tot (mg) is the amount of particulate mass on the particulate filter, $MaPartFilter$ (g) is the particulate filter gas mass, WF is the regulation weighting factor (which depends on the legislation and test cycle) and Flw_EqvCVS (kg/s) is the microtunnel diluted flow and can be calculated by:

$$Flw\_EqvCVS = \frac{Flw\_Exhaust \times MaPartFilter}{MaPartFilter - MaDilAir} \quad (3)$$

where $MaDilAir$ (g) is the dilution air mass and $Flw\_Exhaust$ is the exhaust flow.

As shown, even if the $Spec\_Partcyle$ calculation process is automated, human intervention is still required to weigh the filter.

In such context, machine learning algorithms have been the subject of several analyzes to reduce manual labor and predict parameters. In reference [7], the authors build an artificial neural networks (ANN) based on data of tests of the parameters NOx, HC, CO and smoke. The R^2 value between test and train data are 0.9771, 0.8663, 0.8917 and 0.9858 respectively. In [8], the authors generate a model to predict

NOx using Support Vector Machines. NOx emissions were predicted with a reasonably good accuracy for both training and testing datasets. Furthermore, the squared correlation coefficient of the model was 0.9905 for the training dataset and 0.9768 for the testing dataset. Besides, it was also found that the relative errors of more than 95% for the training data samples and 90% of the test data samples lied within 10%. In reference [9], the authors build an ANN model to predict NOx and reach an estimation error of all test between the error band within ±10% range.

Although there are several works presenting examples of models for predicting engine emissions, most of them do not include PM prediction and, when it is considered, usually is for a specific application and in repeated controlled condition [10, 11].

In a less controlled scenario, results of tests show a high variation of PM measurement. This parameter is affected by several variables, which can make tough to build a robust forecasting model when different variables are considered, for example, different laboratories, engine and fuel. This is what is generally found in industry [12].

In this case, an option would be to build a predict model using random forests algorithm, it is a machine learning technique that builds an ensemble of classification (or regression) trees. With this technique, no precise information is required about the form of the relationship between response and input variables [13].

Besides that, it is a powerful machine learning algorithm. For a better understanding, it is necessary to know how decision tree algorithm works. Decision tree is an approach of a set of predefined conditional rules, which actions and decisions to be made are based on the path (tree) that is being constructed as binary conditions are faced [14].

The decision tree algorithm first splits the training set in two subsets using a single feature k and a threshold t_k, for example, in one subset all flowers that petal length ≤ 2.45 cm (Figure 4). How does it choose k and t_k? It searches for the pair (k, t_k) that produces the purest subsets (weighted by their size). The cost function to be minimized is given by [14]:

$$(k, t_k) = \frac{m_{left}}{m} \times G_{left} + \frac{m_{right}}{m} \times G_{right} \qquad (4)$$

It is named cost function for classification. The G measures the impurity of the left/right subset, the m is the number of instances in the left/right subset.

Once the training set has successfully split in two, the subsets are split using the same logic, then the sub subsets and so on, recursively. It stops recursing once it reaches the maximum depth (defined by the *max_depth* *hyperparameter*), or if it cannot find a split that will reduce impurity [14]. Figure 4 shows an example of a small decision tree.

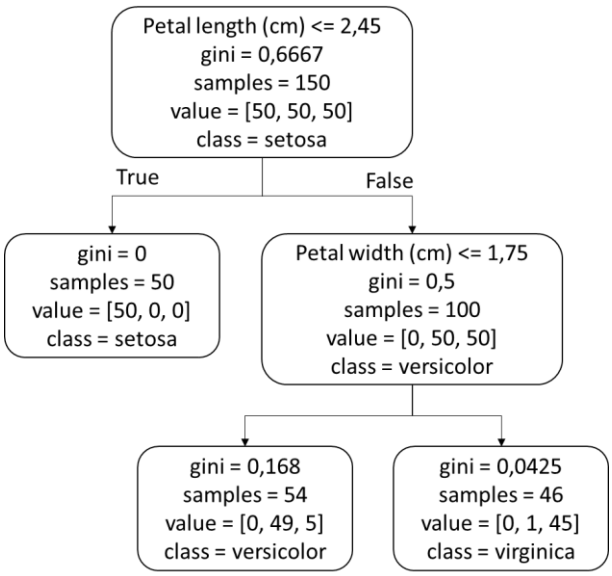

Figure 4 - Decision tree example [14].

A simple way to define the random forests algorithm is a lot of decision trees considered in order to predict a value. For example, a group of decision tree classifiers can be trained, each one in a different random subset of the training set. To make predictions, it is necessary just obtain the predictions of all individual trees, then predict the class that gets the most votes in case of categorical answers, or the mean value of all trees in case of numerical values. Such an ensemble of decision trees is called a random forests [14].

As the majority of machine learning algorithms, random forests have some parameter, called hyperparameters, which should be properly set up to better perform the optimization. Some examples are: the number of trees, maximum tree depth, maximum features and others. The use of incorrect hyperparameter values can generate overfitting or underfitting. In order to find the best hyperparameter combination, the user can run a grid search function to obtain the optimum values for it, therefore the optimum values for the model [14].

PURPOSE – This research aims to develop an automatic way to obtain the parameter $Spec\_Partcycle$ after tests on diesel engines. To that end, random forests algorithm is employed.

MATERIALS AND METHODOLOGY

The research object took place in a truck manufacturer located in Brazil. This company has three engine laboratories with dynamometers and all equipment necessary to carry out tests in accordance with the requirements of EUVI

legislation established in the European Regulation 595/2009 [15].

Several engines of 11, 13 and 16 litters from this truck manufacturer company were included in the analysis. All of them comply with CONAMA Resolution n. 403 [3].

Two tools were developed to perform the analysis: (i) a Visual Basic script to collect data from the files and read the test parameters and (ii) another script, developed in Python, which the main task is to create and run the random forests algorithm. In order to achieve the proposed objective, the methodology followed the steps:

1. Data compiled from all tests performed in test rigs from 2012 to 2019 were collected. Database included results from the three laboratories and 84 parameters were initially considered. Average engine torque, average exhaust flow, air humidity are some examples. The total number of tests in the database was 24000.

2. The following filters were applied to select the database to be used: (i) only engines that comply the CONAMA Resolution n. 403 [3]; (ii) only 11 and 13 litter engine sizes; (iii) only ESC and ETC tests.

3. All testes that $Spec\_Partcycle$ was close to zero, blank or negative value were removed from database.

4. Parameters that had more than 25% blanks data or bad results were eliminated. The remaining parameters were 46 and around 2500 tests results remained as useful data.

5. Parameters that were not representative for the analysis were excluded. Initially, non-numeric parameters were removed. Next, dot plot graphs of "$Spec\_Partcycle$ X Variables" were generated to perform a visual check of the quality of data. The aim of this step was avoiding to input bad data to train the model.

6. Outliers from the remaining variables were excluded.

7. Database was formatted for the reading pattern of Python language. After that, a Python script was executed using all variables left from item 5. The algorithm performs the following tasks:

    a. Scale features to optimize the algorithm. With few exceptions, machine learning algorithms underperform when the input numerical attributes have different scales [14].

    b. Split data in two clusters, ESC and ETC. ESC and ETC are very different in many aspects, due to that, the analysis was performed separately for each test type.

    c. Some tests have parameters that do not have any value associated, the script fill the values with the mean value of the parameter to avoid it´s interference in result.

    d. Split the ESC and ETC clusters in training data (80%) and test data (20%).

    e. Run the training of the random forests and plot the mean square error and the mean error.

    f. Plot the learning curve of the algorithm.

    g. Run grid search function to find the best hyperparameter combination.

8. Parameter importance matrix is then plotted to define the parameters that would be used in the final analysis. Which are: Specific fuel consumption in a cycle (g/kWh), Average CO concentration before aftertreatment system (ppm), Average CO flow before aftertreatment (mg/s), specific CO emissions before aftertreatment system in the cycle (mg/kWh), Average CO_2 flow after the aftertreatment system (mg/s), specific CO_2 emissions after aftertreatment system in the cycle (mg/kWh), Average NOx concentration before aftertreatment system (ppm), specific NO_x emissions before aftertreatment system in the cycle (mg/kWh), smoke opacity (%).

    The parameter importance matrix does not present the variables that have the highest correlation coefficient, but, if one looks at a single decision tree, important features are likely to appear closer to the root of the tree. In other words, these parameters are the ones that most affect the final answer of the algorithm.

9. Python script was executed again using the nine most important parameters and mean error and root mean square error were compared.

10. Hyperparameter grid search was performed to find the optimal parameter combination.

11. Python script was executed again using the optimal parameter combination and mean error and mean square error were again verified.

12. After the random forest model was well trained, new data were collected from new tests that were not part of the initial training set, this new set is named validation set and is composed by 72 new tests (ESC or ETC) performed in engines that follow CONAMA 403 resolution [3], 11, 13 and 16 litter. The 16 litter engines were included in the analysis to increase the range of application of the model and verify the applicability of the model to this engine family too.

    Thereafter, the results obtained by the random forest predict model were compared with those observed from the validation set to check model accuracy.

RESULTS

Before starting the analysis, it was necessary to understand what an acceptable error for the model would be. In this regard, a previous research was performed to understand the reproducibility of $Spec\_Partcycle$ between the three different test cells. Several tests were carried on with the same reference engine, same fuel, same fuel specs but not necessary the same fuel source, and the boundary conditions were slightly different between test cells. The results are shown in Table 2.

Table 2 – PM reproducibility in truck manufacturer laboratories.

Test Type	Number of tests performed	PM reproducibility
ESC	56	14.47%
ETC	57	12.65%

This means that considering only test rigs deviations, the $Spec\_Partcycle$ presented a deviation of ±14.47% for ESC tests and ±12.65% for ETC tests when compared to the mean value.

After the final training of the random forests model, the learning curves of the algorithm were plotted in order to understand if the number of inputs available was enough. Thus, the mean square error (MSE) between random forests model versus test data and random forests model versus training data was plotted, this was done for ESC and ETC models.

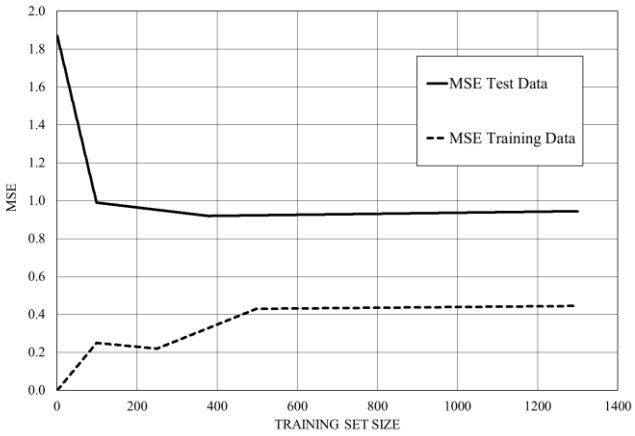

Figure 5 - Learning curves for random forests model - ESC.

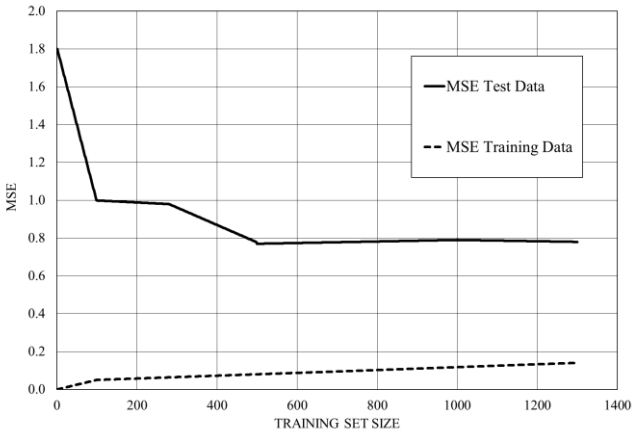

Figure 6 - Learning curves for random forests - ETC.

As represented in Figures 5 and 6, MSE curve is almost flat after 500 samples for ESC and ETC, even though the amount of data available for each test type was around 1250 samples the algorithm was already well trained with 500 samples.

Furthermore, it is possible to notice that MSE is under one for both clusters (ESC and ETC) and, as expected, the MSE for the training data is lower than that for the test data.

ESC RESULTS – Among the tests belonging to the validation set, 42 of them are ESC tests. In Figure 7 is possible to see a comparison between the results of random forests model and the observed results of validation set. The continuous line with dots represents the observed value and the dashed line with triangles is the estimated value by the model. Thus, it is possible to have an idea of the model accuracy. The mean of the observed values is 7.51 mg/kWh.

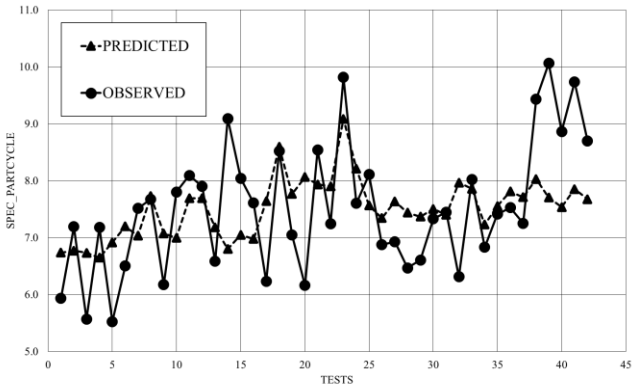

Figure 7 - Predicted and observed *Spec_Partcycle* for ESC.

Figure 8 shows the distribution of the error, it is possible to notice the most of errors are between -1.293 mg/kWh and 1.142 mg/kWh. The mean error between predicted and observed is 0.8 mg/kWh, the error standard deviation is 0.57 mg/kWh. Therefore, if we consider that error follows a normal distribution, 95% of the errors is between ±1.96. This means that for 95% of accuracy the tolerance range will be ±1.96.

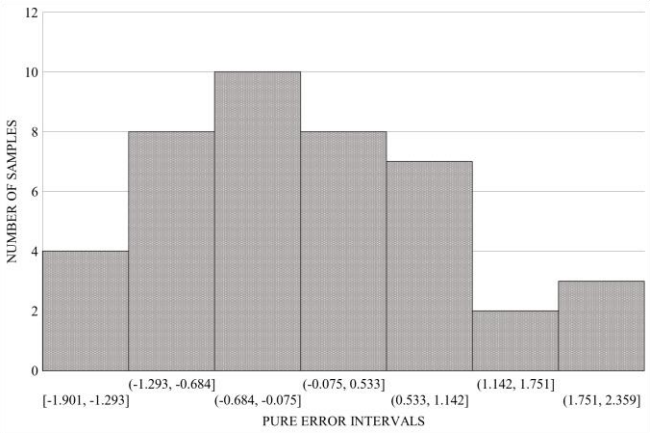

Figure 8 - ESC error distribution.

ETC RESULTS – Among the tests belonging to the validation set, 30 of them are ETC tests. In Figure 9 is possible to see a comparison between the results of random forests model and the results obtained in tests is made. The continuous line with dots represents the observed value and the dashed line with triangles is the estimated value by the model. Also, upper and lower limits considering an interval of confidence is represented. This way is possible to have an idea of the model accuracy. The mean value of the observed results is 15.93 mg/kWh

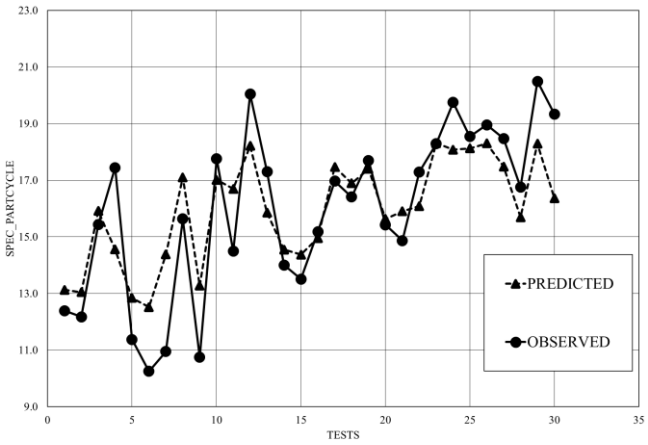

Figure 9 - Predicted and observed *Spec_Partcycle* for ETC.

Figure 10 shows the distribution of the error. It is possible to notice the most of errors are between -1.143 mg/kWh and 1.599 mg/kWh. The mean error between predicted and observed is 1.25 mg/kWh and the standard deviation error is 0.9 mg/kWh. Therefore, if we consider that the error follows a normal distribution, 95% of the error is between ±3. This means that for 95% of accuracy, the tolerance range will be ±3.

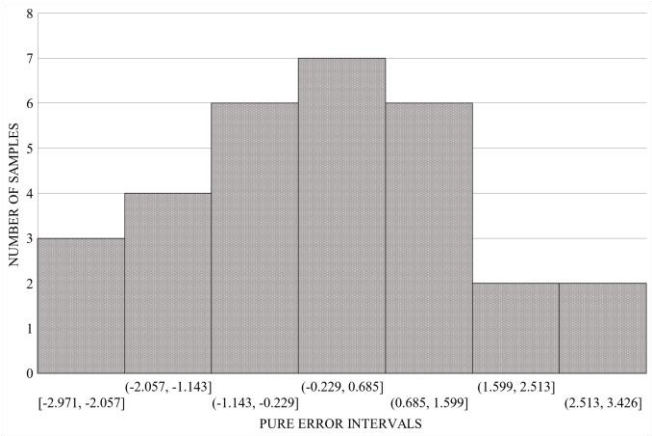

Figure 10 - ETC error distribution.

DISCUSSION

The errors observed in the random forests model are higher than the emissions prediction models built in [7-11]. Several reasons may have contributed to this: the references analyses were carried out in a controlled environment, the same laboratory, the same engine, tests running in sequence, while this research was performed in a wide field, three laboratories were considered, different engines, tests took place on different days.

A fairer comparison to the random forests models generated in this research would be the analysis of

reproducibility that has results represented in Table 2. In this analysis, the boundary conditions are more similar to those used to build the database of this research.

It is also possible to notice that the results in Table 2 has a greater deviation than the reference works [7-11], the reason for this is probably because several items can impact the result of $Spec\_Partcyle$: Exhaust gas flow rate to equipment; engine production of PM (surely not perfectly repeatable); engine service equipment (dyno, cooling water and fuel temperature, combustion air conditioning); power measurement (speed and torque); exhaust gas flow rate measurement (or combustion air flow rate and fuel flow rate); sampling filters conditioning chamber (temperature and humidity stability and cleanliness); micro-balance accuracy and stability, and when adding up more items like: different laboratories, different days, different measuring equipment, for sure a larger deviation will occur.

Considering this scenario, it was considered a more realist target for the model to achieve the same errors shown in Table 2. This means that, if the random forest model is perfect, the minimum error for the ESC tests would be given by:

$$\text{error} = \pm \text{mean observed value in ESC validation data} \times 0.1447 \quad (5)$$

Where the mean observed value in ESC validation data is 7.51. This way, the minimal error is ±1.08 mg/kWh.

Same analysis can be performed for ETC tests, target error is given by:

$$\text{error} = \pm \text{mean observed value in ETC validation data} \times 0.1265 \quad (6)$$

Where the mean observed value in ESC validation data is 15.93. This way, the minimal expected error is ±2.02 mg/kWh.

Errors presented by the random forests model when compared to the observed values in validation data are ±1.96 mg/kWh for ESC and ±3 mg/kWh for ETC tests. Therefore, the difference between the minimal expected error and the error reached is less than 1mg/kWh for ESC and ETC models.

As already mentioned, random forests algorithm is recommended when no precise information about the form of the relationship between response and input variables is available. This is exactly the situation faced in the analyses, and the difference between the minimal expected error and the observed error (less than 1mg/kWh) support this statement.

CONCLUSION

The amount of particulate mass of a test cycle is impacted by different variables. Predicting the $Spec\_PartCycle$ value requires a robust method for correlating the best variables. Considering the results obtained and the corresponding errors, the random forests algorithm seems to be an appropriate method for this analysis.

The $Spec\_PartyCycle$ prediction model does not replace the traditional way of working. It is still be necessary to weigh the filter to obtain the amount of impregnated mass. However, it can be used in cases where no operator can conduct the test or when measuring equipment is unavailable. Another application would be when it is necessary to investigate whether the particulate measurement equipment is working well, comparing the observer results with the prediction of the model.

REFERENCES

[1] FAYYAZBAKHSH, A.; PIROUZFAR, V. Investigating the influence of additives-fuel on diesel engine performance and emissions: Analytical modeling and experimental validation. **Fuel**, ISSN: 0016-2361, v. 171, p. 167–77, 2016.

[2] USLU, S.; CELIK, M.B. Combustion and emission characteristics of isoamyl alcohol-gasoline blends in spark ignition engine. **Fuel**, ISSN: 0016-2361, v. 262, 2020.

[3] BRAZIL. Conselho Nacional do Meio Ambiente. **Resolução CONAMA n. 403**. Brasília – DF. Available in: http://www2.mma.gov.br/port/conama/legiabre.cfm?codlegi=591. Access: 12/12/2019.

[4] EUROPE UNION. Directive 1999/96/ec of the European Parliament and of the Council**. Official Journal of the European Communities,** L44, 1999. Available in: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2000:044:0001:0155:EN:PDF. Access: 05/01/2020.

[5] BURTSCHER, H.; MAJEWSKI W. A.; KHALE, I. A. PM Measurement: Collecting Methods. **DieselNet Technology Guide,** 2016. Available in: https://www.dieselnet.com/tech/measure_pm_col.php. Access: 15/11/2019

[6] BURTSCHER, H. Measurement and characteristics of combustion aerosols with special consideration of photoelectric charging and charging by flame ions. **Journal of Aerosol Science**, ISSN: 0021-8502, v. 23(6), p. 549–595, 1992.

[7] AYDINA, M.; USLUB, S.; ÇELIK, M. B. Performance and emission prediction of a compression ignition engine fueled with biodiesel-diesel blends: A combined application of ANN and RSM based optimization. **Fuel**, ISSN: 0016-2361, v. 269, 2020.

[8] LIU, B.; HU, J.; YAN, F.; TURKSON, R. F.; LIN, F. A novel optimal support vector machine ensemble model for NOX emissions prediction of a diesel engine. **Measurement**, ISSN: 0263-2241, v. 92, p. 183–192, 2016.

[9] KENANOĞLU, R.; BALTACIOĞLU, M. K.; DEMIR, M. H.; ERKINAY ÖZDEMIR, M. Performance & emission analysis of HHO enriched dual-fuelled diesel engine with artificial neural network prediction approaches. **International Journal of Hydrogen Energy,** ISSN: 0360-3199, 2020. In Press.

[10] MAASS, B.; STOBART, R.; DENG, J. Diesel engine emissions prediction using parallel neural networks. In: AMERICAN CONTROL CONFERENCE, 2009. **Annals**…St Louis 2009. p. 1122-27.

[11] YUANWANG, D. An analysis for effect of cetane number on exhaust emissions from engine with the neural network. **Fuel**, ISSN: 0016-2361, v. 81, p. 1963–1970, 2002.

[12] ANDERSSON, J.; MAMAKOS, A.; GIECHASKIEL, B.; CARRIERO, M.; MARTINI, G.; Particle Measurement Programme (PMP) Heavy-duty Inter-laboratory Correlation Exercise (ILCE_HD) Final Report. **Journal Scientic and Technical Reports,** ISSN 1018-5593, 2010

[13] GILLES L. **Understanding Random Forests from theory to practice**. p.221. PhD dissertation. Faculty of Applied Sciences Department of Electrical Engineering & Computer Science, Liège University, Liège, 2014. Available in: https://arxiv.org/pdf/1407.7502.pdf. Access: 30/01/2020.

[14] GÉRON A. **Hands-On Machine Learning with Scikit-Learn and TensorFlow.** 1 ed. USA: O'Reilly Media. 2017.

[15] EUROPE UNION. Regulation (EC) No 595/2009 of the European Parliament and of the Council**. Official Journal of the European Communities,** 2009. Available in: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:188:0001:0013:EN:PDF