Determining predictor variables of HC, CO, and CO2 emissions using decision tree models

A Larasati1, Y W Chen2, A M Hajji3, A F P Mahardika1 and V E B Darmawan1

1 Department of Industrial Engineering, Universitas Negeri Malang, Indonesia
2 Department of Industrial Engineering and Management, Da-Yeh University, Taiwan
3 Department of Civil Engineering, Universitas Negeri Malang, Indonesia

Abstract. HC, CO, and CO2 are three gasses emitted from heavy duty construction equipment, such as backhoe that contain carbon and hugely affect the environment. This research aims to determine predictor variables of HC, CO, and CO2 emissions level using decision tree. The emissions level is categorized into three classes: low, medium, and high. The predictor variables decision tree are related to the backhoe operation and specification, including backhoe type, engine technology tier, RPM, MAP, horsepower, backhoe age, and intake temperature. This study runs 12 cycles for each model to classify HC, CO, and CO2 level. The results indicate that each gas has a different order of the important level for the predictor variable. Decision tree model to classify HC emission level show the top three most important predictor variables are RPM, temperature, and MAP. On the other hand, the top three most important variables to classify CO are backhoe type, RPM, and temperature. The last decision tree model to classify CO2 level show RPM, MAP, and backhoe type as the most important predictor variables.

1. Introduction

Mostly during construction operations, air pollutant emissions are produced from the exhaust of heavy equipment at the construction site [1]. Backhoe is one of heavy equipment used in the construction site. Air emissions produced by the equipment are vary, three of them are Hydrocarbons (HC), Carbon monoxide (CO), and Carbon dioxide (CO2). CO2 or carbon dioxide is one of gas that gives hugely impact on environment, whether it is resulted from bikes, car, or heavy equipment. Mostly, CO2 is considered to be the major contributing component of the global warming [2]. Hydrocarbons (HC) are composed of unburned fuels as a result of lack in temperature. Normally, diesel engines emits low levels of hydrocarbons, because HC is commonly occurred at light loads [3]. CO or carbon monoxide are results from the incomplete combustion. CO is created when the process of oxidation does not occurred perfectly[3]. Even though CO produced during the droplets in diesel engine is big enough, it is minimal chances to occur in diesel engine, especially backhoe. Because of the combustion in diesel engine are lean combustion engines, nonetheless the concentration of HC and CO is minimal.

Mostly carbon dioxide are produced even the backhoe are in idle or non-idle conditions. The emissions value which have been increasing day after day making the company to work on reducing the pollutant emission from heavy equipment. Currently 35% reduction in usage time of backhoe leads to a reduction of approximately 15% in total emissions of backhoe on site construction [4]. Therefore
estimating the emissions gas produced are giving impact on environment around the construction site. And by using predictive model are also can estimating the emissions produced by heavy equipment [5].

Currently, the emissions are estimated mostly from NONROAD model from the EPA, because it was difficult to estimate the emissions from backhoe [1]. That is why this study aims to use data mining with a classification model to estimate the class of the emissions released from backhoe. Decision tree model is one of a technique in data mining which is classified by following the path from the root of the tree to the leaf and it determines which label class it goes [6]. Decision tree is also considered easy to use and to understand. The tree are mostly concise. The large tree with a lot of nodes or branch is less understandable. The output of decision tree are just not a trees that gives nodes and class label that has been explained above, but also an accuracy rate that can be used to measure the accuracy of the tree model [7]. Decision tree can also extract some of the variance variables from the training data and then increases the accuracy of the lower level branches [8]. In brief, decision tree look alike shown in figure 1.

Based on Figure 1 above, every branch has nodes and it contains the classification class. It is why classification are playing important role in decision tree model. If the data isn’t classified then the class label in the trees cannot be explained or probably there will not be any class label in the trees. Classification sometimes is not effective when dealing with numerical data, because it will be has very large number of ways to form the ranges of data and cover it. However, discretization of the numerical data can be a solution to the problem [9]. In this study, HC, CO, and CO\textsubscript{2} are classify into low, middle, high level. It means creating three class can be applied to classify the intention level of the gas emission. Because emissions of backhoe are gases and it units are mass per time, creating three classes are suitable. This research focuses on the estimating emissions gas which is HC, CO, and CO\textsubscript{2} by classifying it classes into three using Decision Tree model.

2. Research Methodology

This study used secondary data about gas emissions produced by backhoe construction equipment such as CO, HC, and CO\textsubscript{2}. The secondary data was measured by a research team from North Carolina State University using portable emissions measurement system (PEMS) [10]. Each of gas produced by backhoe are measured in mass per time (g/s) because of 50% total emissions occurred during the operation time [11]. The dependent variables used are HC, CO, and CO\textsubscript{2} level, while the independent variables used are Normalized MAP (Manifold Absolute Pressure), RPM (Revolutions per Minute), Backhoe type, Backhoe model year (age), Temperature, Engine Technology Tier, and Horsepower. Decision Tree model is used to classify each emissions gas that uses a tree graph decisions, which
concludes the value of dependent variables such as HC, CO, CO$_2$ by given the value of independent variables. The method used in this study briefly shown in Figure 2 where the process flows from data preparation to interpreting the output model.

Figure 2. Flowchart of research method

Data preparation is performed on IBM SPSS Modeler by selecting data that are used in the model, and determining the dependent variable, which is HC, CO, and CO$_2$. Not only selecting variables and determining the target, preparation phase also handling missing value on each variables dataset. Missing value may lead to bias in the model. That is why it is important to not have missing values on the variables, especially on target variables [7]. This study also performs data normalization before finishing data preparation phase. Data of CO$_2$ is still measured by mass per time (g/s) but not for HC and CO that have been measured in g/hr. Thus, this study transforms CO$_2$ into mass per hour (g/hr). Other thing, since the range of HC and CO are much less than CO$_2$, this study do data normalization in order to make these three data set have the same range 0-1.

Data preparation are finished and the next step is binning the data. The dependent variables is classified by binning it into three classes based on equal count. But not all gasses can be classified equal. Because if it classified into 3, it will give 33.33% cases on each classes, and the range will be not evenly distributed. So every dependent variables is classified with different upper limit for every classes that can be shown in table 1.

| Class | HC (gram/hour) | CO (gram/hour) | CO$_2$ (gram/hour) |
|-------|----------------|----------------|-------------------|
| 1     | 6.0            | 13.5           | 2.0               |
| 2     | 12.0           | 29.5           | 4.0               |
| 3     | 182.0          | 2589.0         | 13.3              |

After that split the data into two datasets, training and testing. Each training and testing are contained of 50 points respectively. The next step is to put the model with decision tree. There are a lot Parameter used such as pruning value, minimum records per branch, and number of trials. Pruning value used to make a concise tree. By decreasing the value smaller means create more severe pruning, and accurate tree [12]. Set the minimum records branch also can be used to limit the number of branch splits of the tree, because it can grow exponentially large with the maximum value of it is $2^n$ [8]. Number of trials is a method that allows the model to boost it for improving the accuracy rate by forming the basis for building a subsequent tree that forced to focus on the error of the model before [12]. Basically, it is creating a lot of models to boost the accuracy rate. This model can be said optimal measured by the highest accuracy can get on testing datasets.

In this research, a total of twelve cycles had been running with a different parameter given, from cycle 1 until cycle 12. Every cycle has different pruning severity value, and minimum records per branch
value from the smallest value until the highest value can be set. The smallest value can be given for pruning severity value is 0 while the highest is 100. And for minimum records per branch can be set for smallest is 2 while the highest given in this research is 64. Number of trials this research set is 50, so the model is created for 50 times to boost the accuracy rate. The model was stopped when it reached the lowest point of accuracy obtained in testing datasets. The last step is to see the output which cycle gave the highest accuracy it get.

3. Results and Discussion
By initializing different parameters with total of twelve cycles such as pruning value and minimum records of branch, the accuracy results can be seen in table 2.

| Cycle | Parameter | Accuracy Level |
|-------|-----------|----------------|
|       |           | HC (%) | CO (%) | CO\(_2\) (%) |
| 1     | Pruning Value : 75 | Min branch : 2 | 77.34 | 85.19 | 89.84 |
| 2     | Pruning Value : 35 | Min branch : 10 | 77.10 | 85.03 | 89.79 |
| 3     | Pruning Value : 10 | Min branch : 30 | 75.41 | 84.20 | 89.74 |
| 4     | Pruning Value : 100 | Min branch : 10 | 74   | 82.90 | 89.12 |
| 5     | Pruning Value : 75 | Min branch : 30 | 75.73 | 84.31 | 89.70 |
| 6     | Pruning Value : 100 | Min branch : 30 | 72.75 | 82.35 | 89.25 |
| 7     | Pruning Value : 55 | Min branch : 23 | 75.49 | 84.63 | 89.72 |
| 8     | Pruning Value : 45 | Min branch : 10 | 77.25 | 85.25 | 89.69 |
| 9     | Pruning Value : 0  | Min branch : 40 | 74.96 | 84.53 | 89.58 |
| 10    | Pruning Value : 25 | Min branch : 30 | 75.43 | 84.38 | 89.70 |
| 11    | Pruning Value : 40 | Min branch : 30 | 74.96 | 84.54 | 89.69 |
| 12    | Pruning Value : 75 | Min branch : 64 | 74.16 | 83.67 | 89.27 |

In table 2 above, the accuracy rate given is taken from the output of the testing datasets. Cycle 1 is a simple parameter without changing any default value. The result of the accuracy is that HC gives 77.34%, CO gives 85.19%, and CO\(_2\) gives 89.84%. Given pruning value is 75 and minimum branch is 2, CO\(_2\) and HC simply result in the best accuracy, but CO does not perform at the same level. The highest accuracy of CO classification model is obtained in cycle 8, which gives accuracy of 85.25% with decreasing pruning value into 45 and increasing the minimum branch into 10. Based on the accuracy results, increasing the value of minimum records per branch eventually does not increase the accuracy, except for CO gas classification model. Normally, more complex and size of the branch in tree will give higher accuracy [13]. Because the range of data on each classes in CO data set is much higher than those ones in HC and CO\(_2\) data sets, thus when the number of minimum value increases, the accuracy of CO classification model also increases.
Normalizing datasets also affects the accuracy level of the classification models. Moreover, data normalization ensures each independent variable in the classification model have a fair and equal numerical contribution. Thus, it is important that each class has an equal importance for the classification decision [14]. The existing of non-normal distributed data interferes and lead to bias in the algorithm, as a result it lowers the classification model performance, which is accuracy. Boosting the model may also increases the accuracy of the model. Because the given parameter is 50 for the number of trials, it means the classification model creates 50 times of model being run, which makes the model accuracy increased. The bigger the number of boosting cycle does not guarantee higher accuracy [13].

The other useful result from the classification model is the predictor importance level, which explain how important one variable relatively to the other variables. The predictor importance for each independent variable are shown in Table 3.

| Table 3. Predictor Importance HC, CO, and CO₂ |
|-----------------------------------------------|
| Independent Variable | Relative Importance Level |
| | HC (%) | CO (%) | CO₂ (%) |
| RPM | 66 | 33 | 55 |
| Normal MAP | 8 | 2 | 32 |
| Backhoe type | - | 40 | 9 |
| Horsepower | - | 1 | 3 |
| Age (year) | 8 | - | 1 |
| Engine Tier | 5 | 1 | - |
| Temperature | 13 | 23 | - |

Table 3 implies that most independent variables that gives impact on estimating emissions is RPM for three gasses (HC, CO, & CO₂). RPM simply gives significant correlation with the fuel of diesel engine. MAP or Manifold Absolute pressure also gives highly contributes especially on CO₂, because the ability and size of MAP on different type will affect the emissions produced [15].

The most important results obtained from decision tree is the rules that is used to classify an input. The rules of the best model to classify CO₂, CO, and HC are shown in Table 4, Table 5, and Table 6 respectively. Table 4 indicates that the most important predictor variables within the rules to classify CO₂ at low level and high level are similar. Meanwhile, the rules to classify medium level CO₂ is a bit different since it includes intake temperature as one of the most important predictor variable.

| Table 4. Rules to Predict CO₂ Emissions |
|-----------------------------------------|
| Rules | Number of Rules | Three most important predictor variable within the rule |
| Rules 1 (low level) | 12 | Backhoe type, MAP, RPM |
| Rules 2 (medium level) | 38 | MAP, RPM, intake temperature |
| Rules 3 (high level) | 13 | Backhoe type, MAP, RPM |

Table 5 indicates that each rules contains relatively different predictor variable to classify different level of CO emissions. All rules uses backhoe type and RPM as two important variables. However, the third predictor variable is different for each level. The low level CO emissions rule includes MAP, the medium level rule includes intake temperature, and the high level rule includes horsepower.

| Table 5. Rules to Predict CO Emissions |
|----------------------------------------|
| Rules | Number of Rules | Three most important predictor variable within the rule |
| Rules 1 (low level) | 22 rules | Backhoe type, RPM, MAP |
| Rules 2 (medium level) | 23 rules | Backhoe type, RPM, Intake temperature |
| Rules 3 (high level) | 17 rules | Backhoe type, Horsepower, RPM |
Table 6 indicates that the most important predictor variables within the rules to classify HC at medium level and high level are similar. Meanwhile, the rules to classify low level HC is a bit different since it includes horsepower as one of the most important predictor variable.

| Rules                  | Number of Rules | Three most important predictor variable within the rule          |
|------------------------|-----------------|-----------------------------------------------------------------|
| Rules 1 (low level)    | 37              | Backhoe type, Horsepower, MAP                                   |
| Rules 2 (medium level) | 43              | Backhoe type, MAP, RPM                                         |
| Rules 3 (high level)   | 34              | Backhoe type, MAP, RPM                                         |

Based on the number of rules shown in Table 4, Table 5, and Table 6, it shows that the decision tree to predict HC emissions contains the most rules, which are more than 30 rules for classifying each level. It implies that the decision tree model to predict the class of HC emissions is more complex than the other two models. Furthermore, the rules to classify the medium level of the three gas emissions (CO, CO₂, and HC) are the most complex rules. It explains that the variety combination of data occurs in the medium level gas emissions than in the high or low level.

4. Conclusion
Decision tree can easily model the emissions gas produced by backhoe based on its specification as independent variables. Classification model is used to determine the class of emissions into three classes (low, medium, high), even though not all of the gas emissions model has a good performance (accuracy) on it. CO₂ classification model has the best performance (highest accuracy) because of the range of classes is not as high as CO and HC data sets. Normalization on the data sets also greatly effects on classifying the classes of gas emissions. Normalization effectively increases the classification performance, so the accuracy can be maximized.

Based on the best accuracy model, it conclude that each gas has a different order of the important level for the predictor variables. Furthermore, each class in the decision tree model may not have the same three most important variables in its rule. Decision tree model to classify HC emission level show the order of the predictor variables from the most important to the least important are RPM, temperature, MAP, age and engine technology tier. The three most important variables used in the rule to determine the class of HC emissions are the same for rules 2 and 3, but not for rules 1. The decision tree model to classify CO emission show the order of the predictor variable based on its important level are backhoe type, RPM, temperature, MAP, horsepower and engine tier. The three most important variables used in the rule to determine the class of CO emissions are different for each rules. The last decision tree model to classify CO₂ level show the important rank of the predictor variables are RPM, MAP, backhoe type, horsepower, and engine tier. The three most important variables used in the rule to determine the class of CO₂ emissions are the same for rules 1 and 3, but not for rules 2. Moreover, the independent variable that has the biggest importance to estimate either HC, CO, or CO₂ is RPM.

References
[1] Wang B Z, et al 2018 Assessment and management of air emissions and environmental impacts from the construction industry J. Environ Plan Manag vol 0 no. 0, pp. 1–24.
[2] Lewis P and Rasdorf W 2017 Fuel Use and Pollutant Emissions Taxonomy for Heavy Duty Diesel Construction Equipment J. Manag Eng. vol 33 no. 2
[3] Res A and Keskin A 2014 The pollutant emissions from diesel-engine vehicles and exhaust aftertreatment systems.
[4] Jassim H S H 2017 Predicting Energy Consumption and CO 2 Emissions of Excavators in Earthwork Operations: An Artificial Neural Network Model.
[5] Hajji A M, et al 2017 Combining Off-the-Job Productivity Regression Model with EPA’s NONROAD Model in Estimating CO₂ Emissions from Bulldozer Civ. Eng. Dimens vol. 19 no.
2, pp. 73–78.

[6] Shaun R, et al 2008 *Educational Data Mining 2008 The 1st International Conference on Educational Data Mining*.

[7] Song Y and Lu Y 2015 *Decision tree methods: applications for classification and prediction* vol 27 no 2, pp. 130–135.

[8] Brydon M and Gemino A 2008 *Classification trees and decision-analytic feedforward control: a case study from the video game industry*, pp. 317–342.

[9] Shao J and Tziatzios A 2018 *Data & Knowledge Engineering Mining range associations for classification and characterization Data Knowl. Eng* vol 118 no September 2017, pp. 92–106.

[10] Lewis P, et al 2015 Engine variable impact analysis of fuel use and emissions for heavy-duty diesel maintenance equipment *Transp Res Rec* vol 2482 no. 1, pp. 8–15.

[11] Sepasgozar S M, et al 2019 *Methods for monitoring construction off-road vehicle emissions: a critical review for identifying deficiencies and directions* *Environ Sci. Pollut. Res* vol. 26 no 16, pp 15779–15794.

[12] Brown E C, et al 2003 *National Park vegetation mapping using multitemporal Landsat 7 data and a decision tree classifier* vol 85, pp. 316–327.

[13] Rawal B and Agarwal R 2019 *Improving Accuracy of Classification Based on C4.5 Decision Tree Algorithm Using Big Data Analytics* Springer Singapore.

[14] Singh D and Singh B 2019 Investigating the impact of data normalization on classification performance *Appl. Soft Comput J*, p. 105524.

[15] Park Y, et al 2016 *Control of the air system of a diesel engine using the intake oxygen concentration and the manifold absolute pressure with nitrogen oxide feedback*.

**Acknowledgement**

The authors would like to acknowledge funding support from Universitas Negeri Malang (UM), which has funded this research through PNBP - UM Research Grant 2019.