Classification of air pollutants API Inter-Correlation using decision tree algorithms

O F Althuwaynee¹, A L Balogun ², A Aydda³ and T Gumbo⁴

¹Department of Real Estate Development and Management, Faculty of Applied Sciences, Ankara University, Ankara, Turkey
²Geospatial Analysis and Modelling Research (GAMR) Group, Department of Civil and Environmental Engineering, Universiti Teknologi PETRONAS (UTP), Malaysia
³Department of Geology, Faculty of Sciences, Ibn Zohr University, Agadir, Morocco
⁴Dept. of town and regional planning, University of Johannesburg, Gauteng, South Africa

E-mail: geospatial63@gmail.com

Abstract. The automated classification of ambient air pollutants is an important task in air pollution hazards assessment and life quality research. Faced with various classification algorithms, environmental scientists should select the most appropriate method according to their requirements and data availability. This study describes several types of Decision Tree algorithms for finding the inter-correlation between dominant air pollution index (API) for PM10 percentile values and four other air pollutants such as Sulphur Dioxide (SO₂), Ozone (O₃), Nitrogen Dioxide (NO₂) and Carbon monoxide (CO), in addition to two other meteorological parameters: ambient temperature and humidity, using 22 months records of active air monitoring station in Penang island (northern Malaysia). Classification analysis for the PM10 API was then performed using non-linear Decision Trees within the R programming environment including: Boosted C5.0, Random Forest, PART, and Naive Bayes tree (NBtree). This is in addition to rpart and tree algorithms, which were used to plot the classification trees. The classification performance of the methods is presented and the best classifier in terms of accuracy and processing time was recommended. In R statistical environment, the process of classification by decision tree methods and the classification rules were easy to obtain, while geographic information systems (GIS) software was used for mapping the study area. Furthermore, the results are clear and easy to understand for environmental and geospatial scientists and relevant agencies, which will facilitate the mitigation of air pollution related disasters in the affected communities.

1. Introduction
The recent development of open source languages and software's, and the enormous amount of modeling tools of supportive libraries have made running air quality models at regional and local scales more efficient. R programming environment makes the output and results production very convenient by reducing uncertainty caused by switching between different statistical analysis platforms [1].

Air pollution represents the condition of air pollutants in the atmosphere at high enough concentrations, within serious or above normal ambient levels. It often has a dominant contributory effect on the quality of life [2]. Air pollutant index (API) is defined in terms of the impact of these pollutants on human health [3, 4]. According to the Department of Environment, API limits are Low
pollution (0-50), Moderate (51-100), Unhealthy (101-200), Very unhealthy (201-300), and Hazardous and risky (301-500) [5].

The daily index report comprises some of the major pollution species data ozone (O_3), particulate matter below 10 μm (PM10), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2) [6].

In Malaysia, the API value of PM10 has generally considered a reliable indicator of air quality conditions because it is always higher than other pollutants [5]. These pollutants are produced directly from various emission sources and chemical reactions (table 1).

Many decision tree algorithms have been widely used and validated by several researchers in different fields due to their efficiency in analysis of linear and non-linear data modeling [7, 8]. In environmental research, [9] compared classification, running time and prediction accuracy of numerous decision tree algorithms (C4.5, CART, NBTree, BFTree, LADTree, REPTree, Random Tree, Random Forest, LMT, FT and Decision Stump) using the SO emission levels of industrial pollutants. C4.5 algorithm had the highest accuracy value while the Decision Stump algorithm had the fastest running time.

[10] used a nonparametric data-driven tree-based analysis method CART (classification and regression trees) decision trees to predict maximum surface O_3 concentrations. The results showed moisture content of air plays a limiting role in the maximum surface O_3 concentration.

Using GIS environment tools started with pollutant sources mapping to help in understanding the effective pollutant extent. [11] applied Danish Operational Street Pollution Model (OSPM) for population exposure to traffic air pollution combined with GIS tools to improve assessment of health impacts and in support of risk management from existing administrative databases. [12] used GIS environment for mapping traffic-related air pollution, specifically for NO2 in Amsterdam, Huddersfield, and Prague. The resultant map showed that the GIS-based pollution map provided reliable estimates of concentrations spatially and temporally.

2. Study area
In Malaysia, the major current sources of air pollution arc; mobile sources (contributes 70-75 %), stationary sources (contributes 20-25%), and open burning sources (contributes 3-5%) [6]. [13] found that SO_2, NO_2 and O_3 are the major air pollutants in the Northern region of Malaysia (Pinang Island included), caused mainly due to the combustion process from vehicles and industries as shown in table 1 [13].

The population of Pinang Island reached 1,575,900 in 2010 and increased to 1,611,100 in 2012. The city's growing population is concentrated in urban and industrial areas, increasing the volume of vehicles [14]. The number of industries has also increased significantly, reaching 1,169. These factors contribute to an increase in API values and a decrease in the number of good, environment-friendly days. It also contributes to a critical bias of meteorological parameters [15].

In August 2005, wind movement from Sumatra Island, Indonesia carried open burning particulate matters and aided degradation of air quality, affecting distant locations, including Southern Thailand and Pinang Island (figure 1). This directly increased the PM10 API value by more than 150 API (Figure 2), which is an indication that air pollutant movement may carry significant effects from external sources [13].
Table 1. Main pollutants used in the current study [13].

| Pollutant | Chemical Sources | Reactions | Direct Emissions Sources | Side effects |
|-----------|------------------|-----------|--------------------------|--------------|
| O3        | Carbon monoxide (CO), nitrogen dioxide (NOx), methane (CH4) and volatile organic compound (VOCs). | | Combustion of fuel from vehicles or industries. Anthropogenic, naturally activities | Human health, crop production, ecosystems |
| CO        | | | Incomplete combustion from vehicles, industries and open burning activities | A poisonous gas leads to cardiovascular disease |
| PM10      | | | From vehicles, marine, traffic, non-engine combustion aerosol, soil, dust and burning plantations | Respiratory system, vision reduction |
| SO2       | | | Incomplete combustion process, coal, and fossil fuels | Respiratory system, ecosystem. Cause acid rain |
| NO2       | nitrogen oxide (NOx) through oxidation process | | Incomplete combustion process, vehicles, industries, power plant | Very toxic to human being and the environment, including the formation of acid rain |

Figure 1. Study area showing Pinang Island and road network (Open Street Map).
Methodology

The data was collected from single air monitoring station (CA0038 USM) for 22 months (1st March 2005 to 31st December 2006) (Figure 2). The data record contains hourly API values of PM10, SO2, O3, NO2, and CO, in addition to the other two meteorological parameters; ambient temperature (°C) and humidity (%), in Penang island (Northern Malaysia).

In the current study, we used GIS to map the surroundings sources of pollution and measure the distances to the receptor station.

Four decision trees were tested and compared; Boosted C5.0 (B.C5.0) [16], is an extension of C4.5 that comes with significant improvements in terms of speed (faster), memory usage (less), tree size (smaller), weighting possibility and winnowing to remove insignificant attributes.

Random Forest (RF) [17] is an ensemble learning that uses multiple models for better performance of a single tree model. Moreover, it provides a variable importance option, which is practically useful when working with high number of variables that need to be optimized. It can also be used in place of forward/backward stepwise selection process.

PART [18] rule-based system created through pruned C4.5 tree repeats extracting the rules that help to remove the covered instances by these rules from data.

Naïve Bayes tree (NBTree) [19] is a hybrid algorithm that creates a decision tree with Naïve Bayes classifier and upon the tree’s growth, a naïve Bayes is constructed for each leaf for the data used locally.

Classification and Regression Trees (CART) packages use many algorithms to plot a simplified pruned tree that split attributes based on the sum of squared error values that minimize a loss function. rpart [17] and tree algorithms were used to plot pruned classification trees. This is preferable because the other models have complex tree leaves, which are impractical to include. The following sequential procedures were implemented in this study:

- API was used in place of real values for the pollutants, while the actual values of humidity and ambient temperature were utilized.

Figure 2. (a) Summary plot of air pollutants in CA0038 USM using weekly average of API record and; (b) percentage of total.
• In environmental research, missing data is the first source of uncertainty. Therefore, the data supplier must perform a data imputation process. This study's data has no significant missing records. Hence data imputation was not required.
• The acquired data was prepared and exported to the R environment. The following related packages were also utilized; rpart, Openair, RWeka, random forest, and C50.
• API of PM10 hourly data was converted into 10 percentiles values to focus on revealing the inter-correlation between API of leading pollutants and other pollutants API ranges.
• The first step in decision tree classification is the learning process, where the data is divided into training and testing data, respectively. PM10 API percentiles (1st March 2005 to 31st October 2006) were used as training data to construct a model to produce the classification rules.
• In the second step, training data was used to test the classification accuracy of the models by cross-tabulation technique, and to choose the optimized decision tree types.
• The final stage entailed testing the prediction accuracy of the models using testing data (1st November 2006 to 31st December 2006). In this study, since the temporal scale of data used is limited to 22 months only, we chose two months (10%) of the data for testing purposes.
• Tabulate Confusion matrix that calculates a cross-tabulation of observed and classified classes with associated statistics. And then, the mean value of the ratio between the predicted (training data for modeling accuracy and testing data for prediction accuracy) was found.

4. Empirical result
Table 3 shows the comparison of various decision tree algorithms which were applied to best classification accuracy (descending listed) between 10 percentiles of PM10 API values and API of SO$_2$, O$_3$, NO$_2$, CO in addition to ambient temperature (°C) and humidity (percentage).
Table 3 was constructed based on variable scoring of importance for each decision tree and modeling accuracy found by a confusion matrix. Boosted C5.0 showed higher classification accuracy with more than 83% of the data (12035 nodes) successfully classified, and 54% prediction accuracy. Boosted C5.0 and PART (pruned C4.5) algorithms showed higher efficiency and that agree with findings of [9]. By considering mean absolute error (MAE) and root mean square error (RMSE), we noticed these values are closer to zero, which is the direct index to cross-validation preferable test value (0 to 1).

| Algorithm | CA (%) | PA (%) | RSME | Kappa index | MAE (sec.) | Variables importance (%) |
|-----------|--------|--------|------|-------------|------------|--------------------------|
| B.C5.0    | 83.3   | 54     | -    | -           | S 100      | CO 100, O3 99.7, SO2 100, NO2 99.7, H 99.9, T 99.6 |
| PART      | 76.6   | 48     | 0.17 | 0.63        | M 34       | CO 104, O3 21.6, SO2 14.2, NO2 14.2, H 9, T 9 |
| RF        | 62.6   | 50     | 0.23 | 0.24        | S 43       | CO 13, O3 22, SO2 10.5, NO2 10.5, H 10.5, T 10.5 |
| NBtree    | 53     | 56     | 0.23 | 0.24        | M 24       | CO 13, O3 22, SO2 10.5, NO2 10.5, H 10.5, T 10.5 |

Table 2. Amount of produced pollutants in Metric ton during 2012 in Malaysia.

| Pollutants | Amount (Mt/year) | Sources Vehicles (%) | Power plants (%) | Industries (%) |
|------------|------------------|----------------------|------------------|---------------|
| CO         | 1,873,730        | 94.9                 | 3.8              | 0.5           |
| NO2        | 877,364          | 60                   | 26               | 6             |
| SO2        | 198,519          | 48                   | 7                | 10            |
| PM10       | 6,049            | 76                   | 15               | 4             |
The recommended Kappa index (rater reliability) of observed level of agreement falls between 0.60 to 0.80, indicating a good deal. Therefore, PART algorithm, with 0.62 Kappa index of agreement level, was obtained, which positively confirmed that data collected in the study are correct representations of the variables measured.

All the algorithms agreed that CO is the most correlated pollutants with PM10, that both resources are produced by incomplete combustion (table 1), while second pollutants are SO$_2$ and O$_3$. The significance of these pollutants is fuzzy as the in PART and Random forest (GINI importance measures plot in figure 3, that closely related to the local random forest function that is used to select the best available split) and Naive bias algorithms, and that due to limited data record as well as different resources contribute of each pollutant (table 1).

All the algorithms agreed on the limited roles of humidity and ambient temperature in PM10 occurrence. This is likely due to the characteristic original composition of particle matters that are produced mainly by anthropogenic factors. However, considering the wind direction and wind speed may add more contribution value. NO$_2$ variable shows low inter-correlation importance. This can be described as its original chemical structure that is generated by Oxidation of NO to NO$_2$ in the atmosphere. It is mainly produced by power plants, while vehicles and industries are the major sources of SO$_2$, PM10, and CO (table 1).

**Figure 3.** Mean decreases Gini plot of random forest tree function variable importance.

Figure 4 was produced using the result of Classification and Regression Trees (CART) split attributes based on values that minimize a loss function, such as sum of squared errors. The highest importance scores for variables are CO (48%) and O$_3$ (38%) and others less than 15%.

In Figure 4a, rpart algorithm produced a short tree using 8288 nodes (day) out of 14500 (days represent training data). 80% of pm10 API falls between percentile 20 to 40 with mean value of 33.9 API, while the value of CO API recorded during 22 months to be ranged between 4 to 7 with mean value of 5.8, which is classified as healthy on API scale. Nevertheless, this study aims mainly to find the intercorrelation between PM10 and other pollutants rather than detect the high or low concentration of API.

Thus, if CO API value, which represents the most important variable among the others based on the results of all the classifiers (Table 1), is smaller than 6.5, then it can refer to P20 of PM10 percentile API. Looking at P20 with 70% of the nodes with more focus, this can be represented by O3. If O3 is less than 38, it can refer to P20, else it refers to P30.

If O$_3$ is greater than 6.5 it refers to P30. This value can be classified further by taking the CO value greater than 12, which refers to P40, else CO less than 12 and greater than 6.5 might refer to P30. That can be represented by Humidity higher than or equal to 92 and can be classified as P20, else it refers to P30.
In figure 4b, tree algorithm shows similar classification with rpart algorithms, except the leave of SO\textsubscript{2} that added the range of values less than 5.5 API to be contributed in addition to O\textsubscript{3} less than 37 API, while it excluded the P30 to match the SO\textsubscript{2} values more than 5.5, and that is why SO\textsubscript{2} has controversial role of importance higher or lower than O\textsubscript{3}.

**Figure 4.** Conditional inference tree of PM10 and other pollutants. P10 to P100 refers to API percentile of PM10, (a) using rpart algorithm; (b) using tree algorithm.

5. **Conclusions**

In the current study, we described several types of decision tree algorithms for finding the inter-correlation between dominant air pollution index (API) for PM10 percentile values and other air pollutants SO\textsubscript{2}, O\textsubscript{3}, NO\textsubscript{2} and CO, as well as ambient temperature and humidity in Penang island (northern Malaysia) using 22 months data record.

The study found that CO, SO\textsubscript{2}, and O\textsubscript{3} are the major pollutants significantly inter-correlating with PM10 and contributing to the degradation of air quality in Penang Island. And that occurs mainly due to the combustion process from transportation modes like private vehicles and developing industries. PART and Random forest show the most stable algorithms by introducing clear decisions about the most important variables that inter correlate with PM10 API value.

R programing environment shows reliability in terms of processing and multiple decision trees algorithms availabilities. More data need to be used to understand why NO\textsubscript{2} shows mild importance results as well as using the wind direction and speed will contribute as added value to current findings. Eventually, using decision trees as environ-metric analysis techniques helps to a better understanding of the air quality pattern of main pollutants that contribute to the high API value in the study area.
References

[1] Althuwaynee O F, Musakwa W, Gumbo T and Reis S 2017 Applicability of R statistics in analyzing landslides spatial patterns in Northern Turkey, 2nd Int. Conf. on Knowledge Engineering and Applications (ICKEA) 221-225

[2] Seinfeld J H and Pandis S N 2016 Atmospheric chemistry and physics: from air pollution to climate change (John Wiley & Sons)

[3] Jiang D, Zhang Y, Hu X, Zeng Y, Tan J.and Shao D 2004 Progress in developing an ANN model for air pollution index forecast. Atmos. Environ. 38 7055-7064

[4] Murena, F., 2004. Measuring air quality over large urban areas: development and application of an air pollution index in the urban area of Naples. Atmos. Environ. 38 6195-6202

[5] Awang M B, Jaafar A B, Abdullah A M, Ismail M N, Abdullah R, Johan S and Noor H 2000 Air quality in Malaysia: impacts, management issues and future challenges. Respirology 5 183-196

[6] Afroz R, Hassan M N and Ibrahim N A 2003 Review of air pollution and health impacts in Malaysia. Environ. Res. 92 71-77

[7] Zhang X, Jiang H, Jin J, Xu X and Zhang Q 2012 Analysis of acid rain patterns in northeastern China using a decision tree method. Atmos. Environ. 46 590-596

[8] Zhao M and Li X 2011 An application of spatial decision tree for classification of air pollution index. 9th Int. Conf. on Geoinformatics 1-6

[9] Birant D 2011 Comparison of Decision Tree Algorithms for Predicting Potential Air Pollutant Emissions with Data Mining Models. J. Environ. Inform. 17(1) 46-53

[10] Burrows W R, Benjamin M, Beauchamp S, Lord E R, McCollor D and Thomson B 1995 CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. J. Appl. Meteorol. 34 1848-1862

[11] Jensen S S 1998 Mapping human exposure to traffic air pollution using GIS. J. Hazard. Mater. 61 385-392

[12] Briggs D J, Collins S, Elliott P, Fischer P, Kingham S, Lebret E, Pryl K., Van Reeuwijk H, Smallbone K and Van Der Veen A 1997 Mapping urban air pollution using GIS: a regression-based approach. Int. J. Geogr. Inf. Sci. 11 699-718

[13] Azman S I, Ahmad M A and Armi S M A 2017 Environmetric Study on Air Quality Pattern for Assessment in Northern Region of Peninsular Malaysia. J. Environ. Sci. Technol. 10 186-196

[14] Tan K C, San Lim H, MatJafri M Z and Abdullah K 2010 Landsat data to evaluate urban expansion and determine land use/land cover changes in Penang Island, Malaysia. Environ. Earth Sci. 60 1509-1521

[15] Awang N R, Ramli N A, Mohammed N I and Yahaya A S 2013 Time series evaluation of ozone concentrations in Malaysia based on the location of monitoring stations. International Journal of Engineering and Technology 3(3) 390-394

[16] Evans B and Fisher D 1994 Overcoming process delays with decision tree induction. IEEE Expert 9 60-66

[17] Breiman L, Friedman J, Stone C J and Olshen R A 1984 Classification and regression trees (CRC Press)

[18] Frank E and Witten I H 1998 Generating accurate rule sets without global optimization In J Shavlik (ed.) Machine Learning Proc 15th Int. Conf. (Morgan Kaufmann Publishers San Francisco, CA)

[19] Kohavi R 1996 Scaling Up the Accuracy of Naive-Bayes Classifiers A Decision-Tree Hybrid In KDD 96 202-207