Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Quantification of PM 2.5 bound Polycyclic Aromatic Hydrocarbons (PAHs) and Modelling of Benzo[a]pyrene in the Ambient Air of Automobile Workshops in Benin City

James M. Okuo
University of Benin
Gregory E . Onaiwu (gonaiwu@biu.edu.ng)
Benson Idahosa University

Research Article

Keywords: Modeling, PM2.5, Auto-mechanic workshop, ambient air, GLiM

Posted Date: June 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1727100/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Quantification of $PM_{2.5}$ bound Polycyclic Aromatic Hydrocarbons (PAHs) and Modelling of Benzo[a]pyrene in the Ambient Air of Automobile Workshops in Benin City.

[1]James M. Okuo and [2*]Gregory E. Onaiwu

[1]Environmental Analytical Research Laboratory, Department of Chemistry, University of Benin, Benin City, Nigeria. E-mail: james.okuo@uniben.edu

[2]Department of Physical Sciences, Benson Idahosa University, P.M.B. 1100, Benin City, Edo State, Nigeria. E-mail: gonaiwu@biu.edu.ng

***Corresponding author:**

Gregory E. Onaiwu

[2]Department of Physical Sciences, Benson Idahosa University, P.M.B. 1100, Benin City, Edo State, Nigeria. E-mail: gonaiwu@biu.edu.ng.

Abstract

The activities of artisans conducted regularly in automobile workshops had been observed to generate pollutants that are not limited to particulate matter (PM), Polycyclic Aromatic Hydrocarbons (PAHs). Thus, this research provided data on the quantification of PAHs coupled with the building of a predictive statistical model for the prediction of benzo[a]pyrene (BaP) in Benin City. The City was divided into four zones, namely North West (NW), North East (NE), South East (SE) and South West (SW) and a total of 180 representative samples were collected from artisans' workshops in both wet and dry seasons using Apex2IS Casella standard pump fitted with conical inhalable sampling (CIS) head at a flow rate of 3.5L/min for 8 hours. Meteorological parameters were collected simultaneously with the $PM_{2.5}$. PAHs were extracted and quantified using GC-FID. The annual average concentration of the total PAHs bounded to $PM_{2.5}$ for NW, NE, SE and SW zones were 519.51 (638.78), 109.16 (169.16), 158.89 (178.40) and 77.65 (89.60) ng/$m^3$for both wet and dry season respectively. A generalized linear model (GLiM) was used to develop a predicted model for the prediction of (BaP) air concentrations in the NW zone.

The results of the selected model among the five trained models obtained with data from NW sampling sites are $R^2 = 0.792$ and Adjusted $R^2 = 0.746$ for model 1 with an overall p-value of 0.01. The proposed model established an approximation to estimate Benzo[a]pyrene (BaP) concentrations in the urban automobile workshops' atmospheres with a reasonable accuracy of 60-72%.

Keywords: Modeling, $PM_{2.5}$, Auto-mechanic workshop, ambient air, GLiM

1.0 Introduction

Automobiles constitute one of the primary modes of transportation for conveying people and goods in any country in the world. The usefulness of this mode of transportation is not without some inherent cost. The major cost is the pollution of different magnitude [13]. Researchers across the world had observed that 48 - 80% of total atmospheric pollution comes from major automobile activities. These pollutants are sulphur dioxide, nitrogen oxide, ozone, carbon monoxide, hydrocarbons and particulate matter [21], [24].

International fora such as Earth Summit focused more attention on the reduction of automobile pollution by recommending the use of ethanol and hydrogen as alternative fuels for powering automobiles [44]. The removal of lead from gasoline and the use of electric trains for inter-city transportation had also been proposed [34]. Most of these suggestions have been implemented in some developed countries while the reverse is the case in developing countries.

Automobile activities involved working with and spilling fresh and used oils, greases, petrol, diesel, battery electrolyte, paints, welding electrodes, iron filing machines, and other materials which generate organic (PAHs), inorganic (heavy metals) and particulate matter (PM).

Epidemiological studies of PM particularly $PM_{2.5}$ and ultra-fine particles have been documented [36], [38], [20].

$PM_{2.5}$ generally remains suspended in the atmosphere for a few days and has movement in the range of a few to hundreds of kilometers. $PM_{2.5}$ coupled with the ultra-fine particulate (UFP) may also remain in the atmosphere for a few days to weeks and are most susceptible to fluctuations

in meteorological conditions [10], [40], [49]. The PM is generated primarily from vehicular exhaust, wear and tear of roads, brakes, and tyres, coupled with industrial combustion processes. Their secondary sources include those from agriculture (fertilizers, pesticides, etc.), construction and mining [19, 20], [3], [31]

However, these particulates are usually bounded to polycyclic aromatic hydrocarbons (PAHs) compounds which are a class of complex organic chemicals. The best known PAH is benzo(a)pyrene (BaP), which contains 5 rings. Because of their low vapour pressure, some PAHs are present at ambient temperature in air, both as gas and associated particulates. The lighter PAHs are found almost exclusively in the gas phase whereas the heavier PAHs, such as (BaP), are almost totally adsorbed onto particulate. These compounds are widely distributed in the atmosphere and are among the first atmospheric pollutants to have been identified as suspected carcinogens, particularly (BaP) [35]. USEPA listed 16 PAHs [naphthalene, acenaphthylene, acenaphthene, fluorene, phenanthrene, anthracene, fluoranthene, pyrene, benzo(a)anthracene, chrysene, benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(a)pyrene, dibenzo(a,h)anthracene, benzo(g,h,i)perylene, and indeno(1,2,3-c,d)pyrene] as most priority PAHs to be analyzed in various environmental matrices because of their consistent nature [18]. Of all the priority PAHs, benzo(a)pyrene is termed as the marker or gold standard of the PAHs mixture due to its stability and relatively constant contribution to the carcinogenic activity of particulate-bound PAH as reported by many researchers [23], [41], [48].

The sources of PAHs in the environments, particularly in urban areas, originate mainly from anthropogenic processes, particularly from incomplete combustion of organic fuels at high temperatures. Natural processes, such as volcanic eruptions and forest fires, can also contribute to the ambient existence of PAHs [37].

In Nigeria, almost the entire country has a $PM_{2.5}$ concentration above the WHO guideline of 25μg/m^3 (24-hour mean) and 10μg/m^3 (annual mean) [28]. This presents an environmental health burden arising from the potential risk of continuous exposure to a dangerous level of $PM_{2.5}$

Currently, in Nigeria, there is little or no data on automobile activities emitted fine particulate matter and its associated pollutants. However few researchers have conducted studies on the baseline, spatial and temporary variation of respirable $PM_{2.5}$ and its associated metal level in different cities in Nigeria. Their results showed a comparatively high level of $PM_{2.5}$ that were above the WHO guideline [29], [14], [5], [26], [32], [28], [46].

Also, within the last decade, predictive models have been built for the prediction of total suspended particulate (TSP), but currently, there is no data for fine particulate bound polycyclic aromatic hydrocarbon within the ambient air of automobile workshops in any urban or rural ambient air environment in Nigeria. Thus, the aim of this research is, therefore, to quantify the amount of PAHs and build a predictive model for fine particulate bound Benzo[a]pyrene quantification in the ambient air of automobile workshops using $PM_{2.5}$ and Metrological parameters as the explanatory variables in Benin City.

2.0 Material and Methods

2.0.1 Study area

Benin City, the capital of Edo State, lies between latitude 6º23'55" N to 6º27'39" N and longitude 5º36'18" E to 5º44'30" E [2]. It has a tropical climate, characterized by two distinct seasons, the wet and dry seasons. In this study, the auto-mechanic workshops in Benin City were divided into four zones (NW, NE, SE, and SW) and delineated as shown in Fig.1.

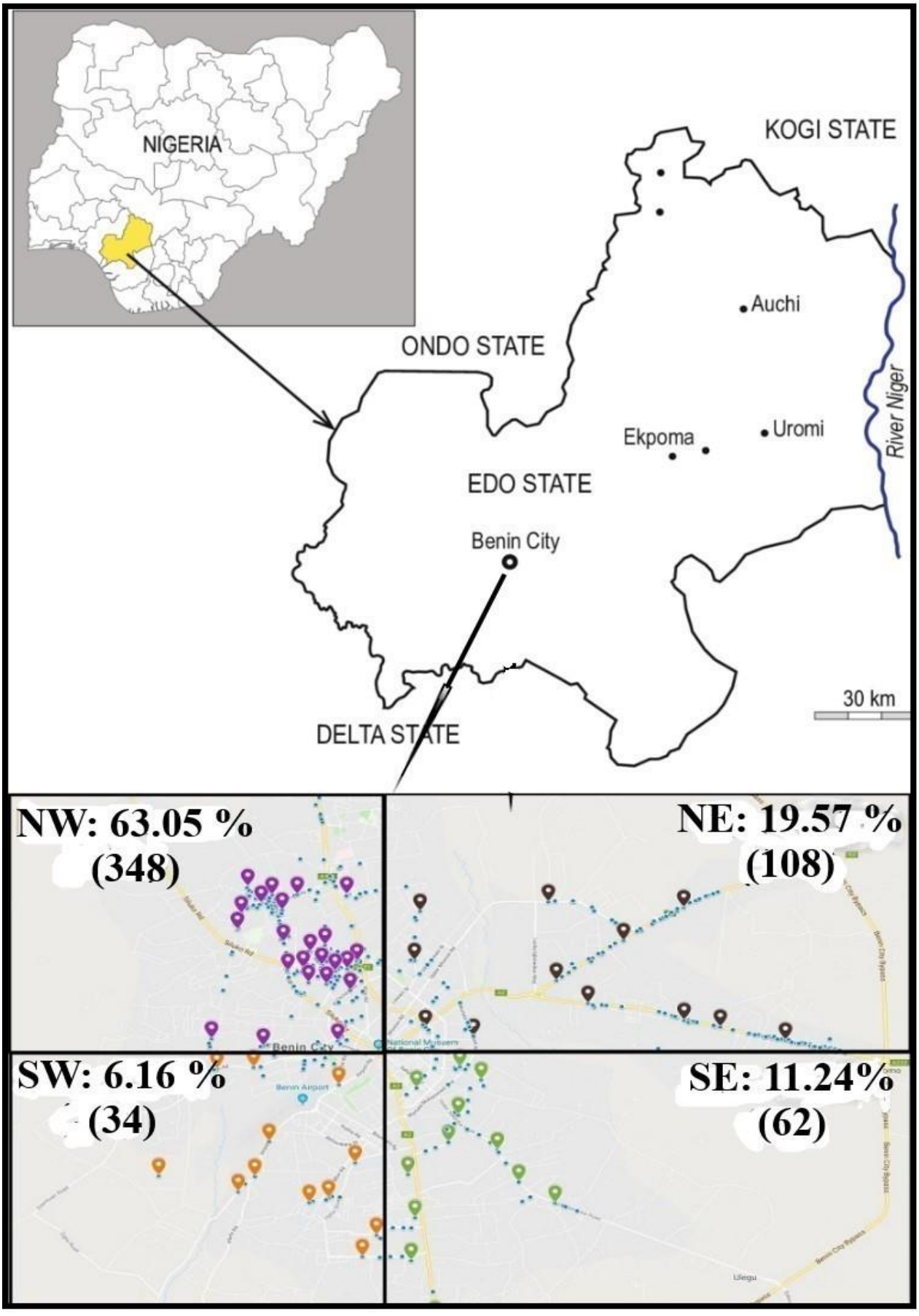

Fig. 1. Map showing the sampling sites of Auto-mechanic workshops in Benin City.

2.0.2 Sampling Strategy for $PM_{2.5}$

A total of 552 auto-workshops were enumerated for this study. Two-thirds (i.e. 368) of the 552 auto-workshops which comprise at least four different artisans such as panel beaters, battery chargers, spray painters, and auto mechanics were selected as sampling sites. However, arising from the similarities in their operations and the clustered nature, the sampling sites were scaled down to 24 in the northwest (NW) and 12 sampling sites each in the Northeast (NE), Southeast (SE) and Southwest (SW) zone as presented in Fig. 1.

Samples were collected three times from each site for one year and a total of 180 air particulates samples were collected. Temperature, relative humidity, pressure, wind speed, wind direction, solar radiation and ultra-violet radiation were also measured and recorded during samplings using standard methods.

The $PM_{2.5}$ samples were collected into 37 mm diameter quartz filters at a height of 1.5 to 2.0 meters above the ground using Apex2IS Casella standard pump, coupled with conical inhalable sampling (CIS) head at a flow rate of 3.5 liters per minute (LPM) for 8 hrs.

A pre and post-field calibration of the pump was carried out for each field sampling to meet the recommended flow rates of ±10% for each sampling period [4]. Before sampling time, the 37 mm diameter quartz filters were treated at 450-500°C for 4 hrs in a muffle furnace while the polyurethane form (PUF) was purged with dichloromethane (DCM) in a soxhlet apparatus at 45°C to eliminate traces of any organic compound that may be on the filter paper. They were thereafter stored in a DCM pre-treated dark plastic bag to prevent photo-oxidation and sealed. The stored (PUF) and quartz filter were equilibrated in a desiccator for 48 hours to eliminate the effect of humidity and also to obtain accurate $PM_{2.5}$ measurements before and after sampling. The pre-weighed filters and the PUF were re-weighed using a four-digit balance with a sensitivity of ±0.1mg and immediately inserted separately into a 100mL brown sampling bottle containing 25mL acetone/dichloromethane (1:1), stored under 4°C for extraction and analysis [45], [7].

2.0.3 Collection of Meteorological Parameters

Meteorological data such as ambient temperature, rainfall, relative humidity (RH), wind speed (WS), wind direction (WD), solar radiation and ultra-violet radiation were recorded through an automatic weather monitoring system (Professional weather station) mounted at 2.5 to 3.0 meters above the ground level at each sampling location closely beside the $PM_{2.5}$ sampler. It was programmed to collect data at an interval of 5 minutes and stored in memory, the recorded measurements were downloaded to a computer using the weatherSmart app [23]

2.0.4 Analytical procedure

2.0.4.1 Sample Extraction, Instrumentation and Analysis

The $PM_{2.5}$ sampled in the quartz filters were ultra-sonicated for 15 mins and allowed to cool for 5 mins [47]. This process was repeated three times. The resultant organic extract was filtered and cleaned through a 0.45 µm filter paper into a column chromatography employing a silica gel column of 7.5 cm in length and 1.5 cm in diameter. To remove any interference of aliphatic hydrocarbons, the column was first eluted with 50 mL of hexane followed by a second elution using DCM. The extract was then concentrated to 5 mL using a rotary evaporator at a temperature of $\leq 40°C$ and allowed to cool to room temperature. It was further reduced to 1 to 1.5mL under a gentle stream of nitrogen gas (N_2) and transfer into a 1 mL brown sampling vial, wrapped in aluminum foil and stored at a temperature below 4ºC for analysis using HP Agilent technology 6890 Gas Chromatography (GC) system equipped with a flame ionization detector. The samples and calibration standards (0.1, 0.2, 0.4, 0.8, 1.6 and 3.2 µg/ml) of 1µL were injected in split/splitless mode (1.5 minutes split time). Polycyclic aromatic hydrocarbon (PAH) analytes separation was carried out on a 30 m low polarity GC column (0.25 mm inner diameter, 0.25 µm film thickness) with the carrier gas; helium (He). The temperature was set at 70°C. This was then ramped to 330°C at a rate of 10°C/min and held for 30 minutes to obtain the relative response value or relative response factor. The concentrations of the analyte were determined from the calibration graph [23], [1]

2.0.5 Modeling

A Log of the generalized linear model was used for the prediction of (BaP) as shown in Equations (1) and (2). Eqn. (1) represents the experimental variables while Eqn. (2) represents the predicted variable. Where α = intercept (a constant value that never changes within a model),

$$\ln \mathrm{Y} = \alpha + \beta_1 \ln X_{i1} + \beta_2 \ln X_{i2} + \beta_3 \ln X_{i1} \ln X_{i2} + \mathrm{K} + \beta_{kj} \ln X_{kj} + \varepsilon \quad (1)$$

$$\ln \hat{\mathrm{Y}} = \alpha + \beta_1 \ln X_{i1} + \beta_2 \ln X_{i2} + \beta_3 \ln X_{i1} \ln X_{i2} + \mathrm{K} + \beta_{kj} \ln X_{kj} \quad (2)$$

$\beta_1, \beta_2, \beta_3 \ldots \beta_k$ = the weight or slope or a regression coefficient (which determines how much weight the independent variable $X_1, X_2, X_3 \ldots X_k$ is contributing to the model),

X_1X_2 = interaction effect of the predictors,

i = 1, 2… n represent sample size,

k and j = the i th contribution of the regression coefficient (β) and the predictors (X) utilized in the experiment.

ln = Natural logarithm

The predicted (BaP) based on the estimation of the results is presented in Eqns. 3, 4, 5, 6 and 7 respectively.

Model 1:

$$\begin{aligned}\ln BaP = {} & 444.8712 - (119.0679 \times \ln T) - (22.9328 \times \ln WS) - (89.4700 \times \ln PM_{2.5}) - \\ & (0.9441 \times \ln UVR) + (25.1369 \times \ln T \times \ln PM_{2.5}) + (2.7951 \times \ln P \times \ln WS \times \ln PM_{2.5}) \\ & - (4.5732 \times \ln T \times \ln WS \times \ln PM_{2.5})\end{aligned} \quad (3)$$

Model 2:

$$\begin{aligned}\ln BaP = {} & 409.8167 - (112.7076 \times \ln T) - (14.9615 \times \ln WS) - (72.491 \times \ln PM_{2.5}) - \\ & (0.3027 \times \ln UVR) + (20.4776 \times \ln T \times \ln PM_{2.5}) + (1.5069 \times \ln P \times \ln WS \times \ln PM_{2.5}) \\ & - (2.3523 \times \ln T \times \ln WS \times \ln PM_{2.5})\end{aligned} \quad (4)$$

Model 3:

$$\ln BaP = 483.5522 - (133.4826 \times \ln T) - (16.8834 \times \ln WS) - (84.0285 \times \ln PM_{2.5}) - (0.3890 \times \ln UVR) + (23.7627 \times \ln T \times \ln PM_{2.5}) - (1.5871 \times \ln P \times \ln WS \times \ln PM_{2.5}) - (2.4296 \times \ln T \times \ln WS \times \ln PM_{2.5}) \quad (5)$$

Model 4:

$$\ln BaP = 386.2221 - (105.8567 \times \ln T) - (18.0034 \times \ln WS) - (70.7210 \times \ln PM_{2.5}) - (0.2275 \times \ln UVR) + (19.8317 \times \ln T \times \ln PM_{2.5}) + (1.8734 \times \ln P \times \ln WS \times \ln PM_{2.5}) - (2.9402 \times \ln T \times \ln WS \times \ln PM_{2.5}) \quad (6)$$

Model 5:

$$\ln BaP = 330.4594 - (89.2878 \times \ln T) - (16.5043 \times \ln WS) - (55.3932 \times \ln PM_{2.5}) - (0.2451 \times \ln UVR) + (15.4613 \times \ln T \times \ln PM_{2.5}) + (1.0644 \times \ln P \times \ln WS \times \ln PM_{2.5}) - (1.4502 \times \ln T \times \ln WS \times \ln PM_{2.5}) \quad (7)$$

T = Temperature, WS = Wind Speed, UVR = Ultra Violet Radiation

The Noise (prediction error) 'ε' which is expected to be very minimal or close to zero for a robust model was evaluated by the following error metrics in Equations 8 to 11 [8, 9].

$$\text{Mean square error } (MSE) = \frac{\sum_{i=1}^{n} \left(\ln y_i - \ln \hat{y}_i \right)^2}{n - v} \quad (8)$$

$$\text{Root mean square error } (RMSE) = \sqrt{\frac{\sum_{i=1}^{n} \left(\ln y_i - \ln \hat{y}_i \right)^2}{n - v}} \quad (9)$$

$$\text{Mean bias error } (MBE) = \frac{\sum_{i=1}^{N} \left(\ln y_i - \ln \hat{y}_i \right)}{n} \quad (10)$$

$$\text{Mean absolute error } (MAE) = \frac{\sum_{i=1}^{n} \left| \ln y_i - \ln \hat{y}_i \right|}{n} \tag{11}$$

Where y_i = Experimental values and $\hat{y}_i$ = Predicted values

2.0.6 Screening for outliers

The Z-scores method was used for screening outliers [39].

$$Z_{scores} = \frac{X_i - \overline{X}}{\sigma} \tag{12}$$

Where σ and $\overline{X}$ are the standard deviation and mean of the distribution X, while X_i is the value of the feature X for the ith sample and the data were transformed using natural Log (i.e. Ln), to fit into Eqn. 2.

2.0.7 Data splitting for training, validation, selection and testing

The five-fold cross-validation approach was used to split the data into training and testing sets as shown in Fig 2.

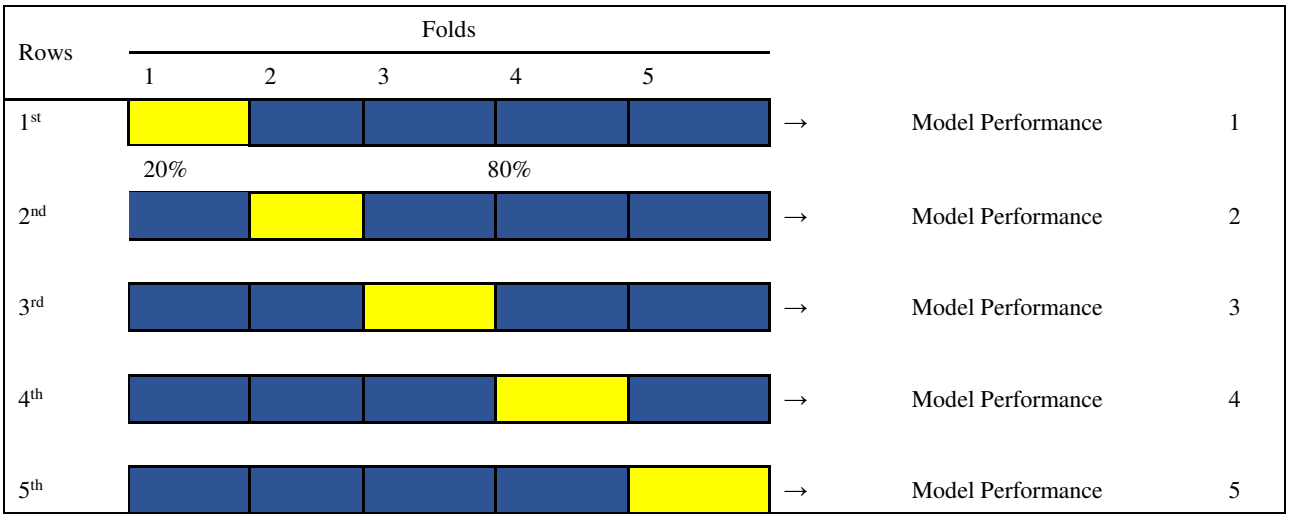

Fig. 2. Illustration of the five-fold cross-validation for model training, testing and selection

The data were split into 'k' parts (in this study, k = 5). One part of the five '5' divisions (i.e. 1/5 or 20%) was used for testing or validation and the remaining K-1 (i.e. 4/5 or 80%) were merged and used for training the model.

2.0.8 Parameter selection approach for training model

The stepwise selection method was used for variable selection by adding or deleting predictors from the existing model based on the F-test statistic and the level of significance ($p>0.05$) using the main effect dialog [9]. The predictors showing a significant correlation with the BaP ($p>0.05$) were retained such that the difference between the R-square and the adjusted R-squared during model training was very minimal. The interaction effect of the predictors was also tested systematically such that their level of significance was less than 0.05 (i.e. p-values ≤ 0.05) as shown in Table 5.

2.0.9 Quality assurance of the model

The quality assurance of the developed/trained model were evaluated based on the satisfaction of the following standard conditional parameters; $R^2_{cal} > 0.6$, Adjusted $R^2_{cal} > 0.6$, $Q^2_{cv(training\ set)} > 0.6$, P (95%) < 0.05, $Q^2_{ext}/R^2_{test}/R^2_{pred} > 0.5$, $RMSE_{test} \approx RMSE_{training}$, $MAE_{test} \approx MAE_{training}$, $(R^2_{training\ set} - Q^2_{cv(training\ set)}) \leq 0.3$, slope: $0.85 \leq K \leq 1.15$, Residual error plot (homoscedastic) and probability plot (plots following a diagonal line) [8, 9].

R^2_{cal} is the coefficient of determination of all independent variables during training,

Adjusted R^2_{cal} is the determination coefficient of the relevant independent variables during training,

$Q^2_{cv\ (training\ set)}$ is the leave many out (LOO) cross-validated for the training set.

2.0.10 Model performance

Model performance was evaluated based on the comparative studies between the training set and the validation/testing set using different error metrics such as MSE, RMSE, MBE and MAE [9]. The predictive power (i.e. performance) of the trained model was also tested using new data set (external cross-validation-Q^2_{ext}) and their slopes (K) evaluated [11], [43], [17], [30].

$$\text{Where; } Q^2_{ext} / R^2_{test} / R^2_{pred} = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i=1}^{n} \left(y_{test} - \hat{y}_{pred(test)} \right)^2}{\sum_{i=1}^{n} \left(y_{test} - \bar{y}_{training\_set} \right)^2} \quad (13)$$

$Q^2_{ext}/R^2_{test}/R^2_{pred}$ is the external validation or predictability of the model and 'k' is the slope of the linear model determined by the experimental variables (predictors) and predicted variable in the external validation-Q^2_{ext}.

3.0 Results and Discussion

The mean concentration of $PM_{2.5}$ bound PAHs and their standard deviation (SD) for NW, NE SE and SW zones are shown in Table 1.

Table 1: Total concentration of the USEPA sixteen PAHs obtained from the NW, NE, SE and SW zone of Benin City.

PAHs component	No. of rings	NW-ZONE		NE-ZONE		SE-ZONE		SW-ZONE	
		Dry season mean ± SD	Wet season mean ± SD	Dry season mean ± SD	Wet season mean ± SD	Dry season mean ± SD	Wet season mean ± SD	Dry season mean ± SD	Wet season mean ± SD
Nap	2	BDL	0.93 ± 0.53	BDL	0.23 ± 0.13	BDL	0.48 ± 0.50	BDL	0.12 ± 0.04
Acy	3	0.95 ± 0.38	1.05 ± 0.62	0.65 ± 0.26	0.13 ± 0.07	BDL	0.13 ± 0.06	BDL	0.24 ± 0.09
Ace	3	BDL	BDL	BDL	0.08 ± 0.05	BDL	BDL	BDL	BDL
Flu	3	BDL	BDL	BDL	0.09 ± 0.03	BDL	BDL	BDL	BDL
Ant	3	4.87 ± 2.94	3.90 ± 2.04	0.42 ± 0.24	0.12 ± 0.04	0.72 ± 0.38	0.06 ± 0.02	0.13 ± 0.05	0.08 ± 0.03
Phen	3	3.28 ± 1.83	3.04 ± 1.68	0.78 ± 0.49	BDL	2.06 ± 1.25	0.48 ± 0.26	0.39 ± 0.15	0.36 ± 0.14
Flt	4	1.55 ± 0.91	1.36 ± 0.71	0.94 ± 0.79	BDL	1.10 ± 0.56	0.77 ± 0.45	1.23 ± 0.95	0.74 ± 0.69
Pyr	4	3.46 ± 2.12	1.17 ± 0.84	1.68 ± 1.38	0.81 ± 0.63	1.85 ± 1.25	1.38 ± 0.99	0.94 ± 0.74	0.72 ± 0.54
B(a)A	4	5.69 ± 1.87	2.96 ± 1.71	2.64 ± 2.07	1.64 ± 1.07	2.01 ± 1.21	2.01 ± 1.43	2.16 ± 1.70	0.85 ± 0.68
Chry	4	9.38 ± 7.62	7.20 ± 5.66	2.61 ± 1.98	3.32 ± 3.03	2.55 ± 1.89	2.13 ± 1.58	1.55 ± 1.10	1.82 ± 1.92
B(k)F	5	14.84 ± 6.69	6.49 ± 4.36	4.73 ± 2.01	3.56 ± 2.23	4.48 ± 3.65	2.43 ± 1.56	2.88 ± 2.03	2.06 ± 1.01
B(b)F	5	52.62 ± 40.87	32.34 ± 28.79	8.4 ± 4.17	11.89 ± 13.45	5.79 ± 3.56	6.24 ± 4.49	3.45 ± 2.14	2.71 ± 1.59
B(a)P	5	73.72 ± 56.80	66.61 ± 40.21	27.56 ± 32.15	12.90 ± 16.36	16.13 ± 10.81	11.75 ± 9.35	12.45 ± 10.47	3.55 ± 1.41
InP	6	96.95 ± 71.25	70.68 ± 39.51	25.03 ±19.32	15.02 ± 7.44	24.98 ± 14.80	23.73 ± 10.35	10.31 ± 5.89	11.18 ± 5.80
D(a,h)A	5	63.96 ± 25.40	134.22 ± 70.74	22.96 ± 15.13	16.24 ± 9.46	35.88 ± 14.80	34.43 ± 16.24	13.13 ± 8.94	7.30 ± 5.68
B(g,h,i)P	6	307.51 ± 157.66	187.56 ± 97.79	70.76 ± 30.49	43.1 ± 20.38	80.85± 37.78	72.87 ± 42.05	40.98 ± 20.50	45.92 ± 25.41
Total ng/m^3		638.78±349.93	519.51±280.43	169.16 ± 67.08	109.13 ± 47.86	178.40 ± 78.68	158.89 ± 80.89	89.60 ± 39.13	77.65 ± 38.94

BDL = Below Detection Limit

The annual average concentration of the total PAHs in $PM_{2.5}$ for NW, NE, SE and SW zones ranged from 280.43 to 988.71, 47.86 to 236.24, 78.68 to 257.08 and 38.94 to 128.73 ng/m^3 respectively. The average concentration of individual PAHs in $PM_{2.5}$ for the NW zone varied from 0.95 ± 0.38 (Acy) to 307.51 ± 157.66 (B(g,h,i)P) ng/m^3 in the dry season and from 0.93 ± 0.53 (Nap) to 187.56 ± 97.79 (B(g,h,i)P) ng/m^3 in the wet seasons. In the NE, the variation in the dry season were 0.65 ± 0.26 (Acy) to 70.76 ± 30.49 (B(g,h,i)P) ng/m^3 and in the wet season 0.08 ± 0.05 (Ace) to 43.1 ± 20.38 (B(g,h,i)P) ng/m^3 were recorded. The recorded average concentration of the individual PAHs in the SE zone were from 0.72 ± 0.38 (Ant) to 80.85 ± 37.78 (B(g,h,i)P) ng/m^3 during the dry season and 0.13 ± 0.05 (Ant) to 40.98± 20.50 (B(g,h,i)P) ng/m^3 in the wet season while for SW zones the values varies from 0.08 ± 0.50 (NaP) to 72.87± 42.05 (B(g,h,i)P) and 0.12 ± 0.04 (Nap) to 45.92 ± 25.41 (B(g,h,i)P) ng/m^3 for the dry and wet season respectively. The mean values of BaP for the NW and NE zones in the dry and wet seasons were 73.72 ± 56.80 (66.61 ± 40.21) and 27.56 ± 32.15 (27.56 ± 32.15). While for SE and SW zones it was 16.13 ± 10.81 (11.75 ± 9.35) and 12.45 ± 10.47 (3.55 ± 1.41) ng/m^3 respectively.

All the obtained mean values exceeded 1 ng/m^3 recommended by the NAAQS and WHO [33]. It was also observed that the concentrations of the high molecular weight (HMW, 4-6 ring) PAHs were significantly higher than the low molecular weight (LMW: 2-3 ring) PAHs across the sampling zones. The percentage contribution of the LMW PAHs across the four zones in the dry and wet seasons were 1.42 (1.72), 1.09 (0.60), 1.56 (0.72) and 0.58 (1.03) % for the NW, NE, SE and SW respectively. While that of the HMW was also revealed to be 98.58 (98.28), 98.91 (99.40), 98.44 (99.28) and 99.42 (98.97) % respectively. This result agrees with the studies carried out by Tan and others [48], which state that the HMW PAHs associates mainly with the fine part of particulate matter (PM) while LMW exists mainly in the coarse part of PM.

3.0.1 Outliers

The Z-score values for the dependent and independent variables as shown in Table 2 were used for screening outliers.

Table 2. The minimum and maximum Z-scores values for the variables used for modeling.

Data variables	n	Z-scores	
		Minimum	Maximum
Temperature	72	-2.1633	1.9056
Relative Humidity	72	-1.708	2.608
Pressure	72	-1.5211	1.7207
Wind Speed	72	-1.834	1.8299
Wind direction	72	-2.3142	2.0535
Ultra-Violent radiation	72	-2.2294	2.0712
Solar Radiation	72	-2.0782	2.2184
$PM_{2.5}$	72	-1.108	2.443
Benzo(a)pyrene	72	-1.3643	1.1167

There are several existing approaches for detecting outliers in a univariate set of data [25]. It was based on the property of the Gaussian distribution curve that 99.5% of the data should lie between 3 and -3. The positive z-score showed that the raw scores are higher than the mean average of zero while the negative z-score reveals that the raw scores are lower than the average mean. However, the z-score values for all the data variables were within the recommended tolerance limit of +3 to -3 as shown in Table 2.

The results of the original and natural logarithmic transformed data for model training are presented in Table 3. In real life, it is often very difficult for most data to follow a normal distribution curve. They are either skewed to the right or the left and as such make the result of the statistical analysis invalid, for not meeting certain statistical conditions such as equal distribution about the mean as often prescribed for linear modeling [16].

Thus the logarithmic transformation in this study was used to minimize the skewness of the original data sets in Table 3. To interpret the targeted output variable (predicted) at the end, we re-transformed it to the original data by exponentiating (exp) the output (X: BaP) with a base of e (i.e. Euler's constant: 2.718) [15].

Table 3. Mean and standard deviation of Parameters used for modeling

	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}	X_{i6}	X_{i7}	X_{i8}	X_{i9}
S/N	T	RH	P	WS	WD	UVR	SR	$PM_{2.5}$	B(a)P
1	30.33 ± 0.32	76.13 ± 1.38	749.87 ± 0.91	6.12 ± 0.48	197.29 ± 20.03	554.12 ± 149.13	929.54 ± 71.57	597.22 ± 41.95	108.26 ± 2.77
2	31.03 ± 0.22	72.92 ± 0.88	749.23 ± 0.34	5.65 ± 0.16	181.71 ± 5.84	593.20 ± 11.78	974.88 ± 60.43	536.11 ± 12.55	110.85 ± 1.48
3	31.75 ± 0.76	70.42 ± 3.83	748.45 ± 0.34	5.63 ± 0.49	181.67 ± 9.55	709.31 ± 106.65	1073.63 ± 62.74	466.27 ± 18.15	109.52 ± 3.44
4	32.37 ± 0.64	68.55 ± 0.63	748.19 ± 0.13	5.21 ± 0.23	200.67 ± 20.12	775.18 ± 73.71	1174.67 ± 21.67	353.74 ± 22.56	106.86 ± 8.27
5	30.53 ± 0.73	74.71 ± 1.38	749.72 ± 0.66	5.85 ± 0.32	178.33 ± 34.76	534.60 ± 109.78	804.99 ± 175.87	321.42 ± 30.92	69.54 ± 23.92
6	29.67 ± 0.28	74.55 ± 2.50	750.63 ± 0.78	6.16 ± 1.04	178.84 ± 54.71	541.47 ± 41.98	874.64 ± 319.06	285.71 ± 15.79	22.06 ± 13.09
7	29.83 ± 0.61	72.54 ± 2.64	752.01 ± 2.50	6.11 ± 1.19	178.13 ± 61.55	496.60 ± 135.61	906.58 ± 301.27	262.46 ± 42.65	22.93 ± 19.61
8	29.21 ± 0.09	70.58 ± 1.87	753.92 ± 1.81	5.13 ± 0.54	155.28 ± 29.44	574.59 ± 41.18	794.92 ± 40.63	193.68 ± 66.80	17.37 ± 9.57
9	27.81 ± 0.51	79.1 ± 4.32	755.08 ± 0.54	4.20 ± 1.08	129.62 ± 48.86	422.76 ± 90.52	774.60 ± 241.24	113.34 ± 16.39	11.90 ± 5.93
10	28.85 ± 1.09	72.85 ± 6.42	754.60 ± 0.35	4.55 ± 1.31	185.25 ± 48.16	552.33 ± 221.62	882.96 ± 166.94	123.12 ± 69.61	14.15 ± 9.62
11	29.68 ± 0.78	70.89 ± 6.43	753.32 ± 1.44	3.20 ± 0.81	177.44 ± 67.24	513.62 ± 55.63	805.73 ± 228.52	93.50 ± 20.85	31.68 ± 12.44
12	29.13 ± 2.28	75.62 ± 10.69	751.18 ± 2.80	4.12 ± 0.03	168.25 ± 29.31	597.21 ± 145.85	935.18 ± 279.47	50.74 ± 9.79	7.63 ± 1.83
13	27.66 ± 1.89	78.54 ± 7.80	752.84 ± 0.58	2.77 ± 0.11	189.74 ± 49.09	431.82 ± 122.02	812.26 ± 255.23	40.66 ± 9.90	116.47 ± 12.04
14	29.15 ± 1.74	71.75 ± 7.23	752.45 ± 0.40	3.61 ± 0.73	207.09 ± 17.61	604.71 ± 116.21	1213.68 ± 230.14	50.64 ± 19.54	104.61 ± 3.74
15	32.64 ± 0.17	62.74 ± 0.81	752.6 ± 0.67	2.57 ± 0.47	162.52 ± 35.38	612.29 ± 168.27	836.37 ± 204.41	78.65 ± 22.66	95.88 ± 25.33
16	31.62 ± 0.47	67.87 ± 1.68	751.75 ± 0.97	3.98 ± 2.05	172.78 ± 77.88	639.66 ± 83.48	1085.79 ± 215.32	244.33 ± 66.24	72.48 ± 19.60
17	30.68 ± 0.44	74.52 ± 1.43	749.55 ± 0.66	5.89 ± 0.39	189.50 ± 5.70	573.66 ± 113.23	952.21 ± 54.28	566.66 ± 43.68	109.56 ± 2.09
18	31.88 ± 0.51	69.48 ± 2.71	748.32 ± 0.24	5.42 ± 0.23	191.17 ± 16.86	742.25 ± 44.61	1124.15 ± 44.98	409.72 ± 50.09	108.19 ± 5.85
19	29.93 ± 0.30	74.63 ± 1.13	750.18 ± 0.86	6.01 ± 0.58	178.59 ± 9.49	538.04 ± 35.49	839.82 ± 243.84	303.57 ± 24.61	45.80 ± 33.53
20	29.49 ± 0.51	71.56 ± 2.55	752.97 ± 2.22	5.62 ± 0.98	166.70 ± 40.54	535.60 ± 66.89	850.75 ± 90.52	228.07 ± 34.05	20.15 ± 7.74
21	28.33 ± 0.79	75.98 ± 5.23	754.84 ± 0.26	4.37 ± 1.16	157.44 ± 7.22	487.55 ± 132.29	828.78 ± 164.98	118.23 ± 35.25	13.03 ± 4.18
22	29.41 ± 0.45	73.26 ± 0.42	752.25 ± 0.28	3.66 ± 0.52	172.85 ± 5.67	555.42 ± 82.10	870.46 ± 185.75	72.12 ± 22.31	29.66 ± 7.35
23	28.57 ± 0.36	74.98 ± 1.39	752.64 ± 0.39	3.19 ± 0.56	198.41 ± 27.23	518.27 ± 57.32	1012.97 ± 146.37	45.65 ± 13.27	110.54 ± 10.51
24	32.13 ± 0.59	65.31 ± 2.91	752.18 ± 0.58	3.28 ± 1.12	167.65 ± 15.25	625.98 ± 115.91	961.08 ± 240.07	161.49 ± 24.05	84.18 ± 21.86

Table 3. Mean and standard deviation of Parameters used for modeling (continued)

	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}	X_{i6}	X_{i7}	X_{i8}	X_{i9}
S/N	lnT	lnRH	lnP	lnWS	lnWD	lnUVR	lnSR	$lnPM_{2.5}$	lnBaP
1	3.41 ± 0.01	4.33 ± 0.02	6.62 ± 0.00	1.81 ± 0.08	5.28 ± 0.10	6.29 ± 0.28	6.83 ± 0.08	8.69 ± 0.07	4.69 ± 0.03
2	3.44 ± 0.01	4.29 ± 0.01	6.62 ± 0.00	1.73 ± 0.03	5.20 ± 0.03	6.37 ± 0.20	6.88 ± 0.07	8.59 ± 0.03	4.71 ± 0.01
3	3.46 ± 0.02	4.25 ± 0.05	6.62 ± 0.00	1.73 ± 0.09	5.20 ± 0.05	6.55 ± 0.16	6.98 ± 0.06	8.45 ± 0.04	4.69 ± 0.03
4	3.46 ± 0.01	4.23 ± 0.01	6.62 ± 0.00	1.65 ± 0.05	5.30 ± 0.10	6.65 ± 0.10	7.07 ± 0.02	8.12 ± 0.07	4.67 ± 0.08
5	3.41 ± 0.01	4.31 ± 0.02	6.62 ± 0.00	1.76 ± 0.06	5.17 ± 0.19	6.26 ± 0.22	6.67 ± 0.22	8.07 ± 0.09	4.20 ± 0.36
6	3.39 ± 0.01	4.31 ± 0.04	6.62 ± 0.00	1.81 ± 0.17	5.15 ± 0.34	6.29 ± 0.08	6.73 ± 0.34	7.96 ± 0.06	2.95 ± 0.68
7	3.39 ± 0.02	4.28 ± 0.04	6.62 ± 0.01	1.80 ± 0.20	5.14 ± 0.40	6.18 ± 0.26	6.77 ± 0.32	7.86 ± 0.16	2.91 ± 0.80
8	3.38 ± 0.01	4.26 ± 0.02	6.63 ± 0.01	1.63 ± 0.11	5.03 ± 0.20	6.35 ± 0.07	6.68 ± 0.06	7.53 ± 0.37	2.71 ± 0.69
9	3.33 ± 0.02	4.37 ± 0.06	6.63 ± 0.00	1.41 ± 0.25	4.81 ± 0.38	6.04 ± 0.21	6.62 ± 0.30	7.03 ± 0.14	2.39 ± 0.52
10	3.36 ± 0.03	4.28 ± 0.09	6.63 ± 0.00	1.49 ± 0.28	5.20 ± 0.28	6.27 ± 0.37	6.77 ± 0.18	7.02 ± 0.53	2.51 ± 0.62
11	3.39 ± 0.03	4.26 ± 0.09	6.62 ± 0.01	1.14 ± 0.27	5.12 ± 0.44	6.24 ± 0.11	6.67 ± 0.27	6.82 ± 0.24	3.41 ± 0.40
12	3.37 ± 0.08	4.32 ± 0.15	6.62 ± 0.01	1.42 ± 0.01	5.12 ± 0.17	6.37 ± 0.23	6.81 ± 0.29	6.22 ± 0.19	2.67 ± 1.36
13	3.32 ± 0.06	4.36 ± 0.10	6.62 ± 0.00	1.02 ± 0.04	5.22 ± 0.27	6.04 ± 0.27	6.67 ± 0.30	5.99 ± 0.25	4.75 ± 0.10
14	3.38 ± 0.06	4.27 ± 0.10	6.62 ± 0.00	1.27 ± 0.19	5.33 ± 0.09	6.39 ± 0.21	7.09 ± 0.21	6.17 ± 0.43	4.65 ± 0.04
15	3.49 ± 0.01	4.14 ± 0.01	6.62 ± 0.00	0.93 ± 0.19	5.07 ± 0.23	6.39 ± 0.26	6.71 ± 0.23	6.64 ± 0.27	4.54 ± 0.28
16	3.46 ± 0.02	4.22 ± 0.02	6.62 ± 0.00	1.30 ± 0.48	5.09 ± 0.44	6.45 ± 0.13	6.71 ± 0.62	8.02 ± 0.62	4.25 ± 0.29
17	3.42 ± 0.02	4.31 ± 0.02	6.62 ± 0.00	1.77 ± 0.07	5.24 ± 0.03	6.33 ± 0.20	6.86 ± 0.06	8.64 ± 0.07	4.70 ± 0.02
18	3.46 ± 0.02	4.24 ± 0.04	6.62 ± 0.00	1.69 ± 0.04	5.25 ± 0.08	6.60 ± 0.07	7.03 ± 0.04	8.31 ± 0.12	4.68 ± 0.05
19	3.40 ± 0.01	4.31 ± 0.01	6.62 ± 0.00	1.79 ± 0.09	5.31 ± 0.08	6.267 ± 0.07	6.74 ± 0.16	7.96 ± 0.17	3.49 ± 0.80
20	3.36 ± 0.03	4.30 ± 0.09	6.63 ± 0.00	1.51 ± 0.20	4.97 ± 0.04	6.14 ± 0.23	6.61 ± 0.14	7.46 ± 0.38	2.46 ± 0.29
21	3.37 ± 0.02	4.29 ± 0.03	6.63 ± 0.01	1.39 ± 0.32	5.02 ± 0.08	6.28 ± 0.14	6.80 ± 0.04	6.92 ± 0.33	2.81 ± 0.36
22	3.37 ± 0.03	4.30 ± 0.03	6.62 ± 0.00	1.25 ± 0.21	5.14 ± 0.04	6.24 ± 0.19	6.72 ± 0.21	6.26 ± 0.39	3.57 ± 1.07
23	3.40 ± 0.08	4.25 ± 0.10	6.62 ± 0.00	1.09 ± 0.24	5.25 ± 0.19	6.27 ± 0.06	6.83 ± 0.23	6.26 ± 0.34	4.67 ± 0.02
24	3.46 ± 0.01	4.21 ± 0.02	6.62 ± 0.00	1.31 ± 0.20	5.11 ± 0.03	6.48 ± 0.11	6.68 ± 0.36	7.94 ± 0.66	4.24 ± 0.04

ln = Natural logarithm, T = Temperature, RH = Relative humidity, P = Pressure, WS = Wind speed
WD = wind direction, UVR = Ultra-violet radiation, SR = Solar radiation, $PM_{2.5}$ = Fine Particular matter, B(a)P = Benzo (a) pyrene.

3.0.2 Model selection

3.0.2.1Internal validation

Table 4 revealed the error evaluation of the five different models and their results. Model 1 satisfied all the necessary conditions among the five trained models. It had the highest Pearson correlation coefficient, R-squared and adjusted R-squared values of not less than 0.74 (74%). When compared to models 2, 3, 4, and 5 their adjusted R-squared values were 0.57, 0.67, 0.60 and 0.55 respectively.

The statistically significant variables (p-value ≤ 0.01) in all the models represented in Table 5 were temperature, wind speed and $PM_{2.5}$ the only exception observed in model 5 was $PM_{2.5}$ have a significance level of $p \leq 0.05$. The interaction effect between these variables was also shown to be statistically significant at a p-value ≤ 0.05. Considering the total mean variable contribution of all the variables, model 1 was statistically significant at p-value = 0.017, coupled with the mean variable observed power computed at $p \leq 0.05$ to be 0.824 (82%). This, therefore, suggested that the actual predicting power of all the predictors for model 1 was about 82% compared to the 100% (1.000) (Table 4). While the actual observed power of the variables and their level of significance for model 2, 3, 4 and 5 were 67% (p-value = 0.120); 74% (p-value = 0.092); 78% (p-value = 0.109) and 49% (p-value = 0.199) respectively. Table 4, also showed the results of some universally accepted error metrics usually used for forecasting (predicting) models.

Table 4. The training and evaluation of five different model results.

Model evaluation	MBE	MAE	RMSE	MSE	R	R^2	Adjusted R^2	Predictive R^2 or Q^2_{cv}
Trained model 1 (80% of data)	0.0000	0.4005	0.4838	0.2340	0.890	0.792	0.746	0.791
Error estimate for 20% new data	-0.4425	0.5525	0.6063	0.3676	0.691	0.477	-	0.601
Average error estimate for 20% resampled data	-0.0050	0.3863	0.4476	0.2229	0.603	0.376	-	0.722
Trained model 2 (80% of data)	0.0005	0.4770	0.6065	0.3678	0.804	0.648	0.570	0.647
Error estimate for 20% new data	0.3438	0.3438	0.4090	0.1673	0.416	0.173	-	0.860
Average error estimate for 20% resampled data	0.0153	0.4572	0.5795	0.3413	0.550	0.337	-	0.635
Trained model 3 (80% of data)	0.0002	0.4253	0.5489	0.3013	0.857	0.734	0.676	0.733
Error estimate for 20% new data	0.0487	0.5988	0.7419	0.5505	0.769	0.592	-	0.117
Average error estimate for 20% resampled data	0.0209	0.3922	0.4897	0.2625	0.456	0.229	-	0.698
Trained model 4 (80% of data)	-0.0005	0.3970	0.5480	0.3003	0.820	0.673	0.602	0.672
Error estimate for 20% new data	-0.5250	0.6400	0.7514	0.5646	0.385	0.148	-	0.695
Average error estimate for 20% resampled data	-0.1018	0.4133	0.5292	0.3168	0.436	0.243	-	0.312
Trained model 5 (80% of data)	-0.0002	0.4088	0.5649	0.3191	0.794	0.631	0.550	0.630
Error estimate for 20% new data	-0.5175	0.6025	0.7258	0.5267	0.429	0.184	-	0.761
Average error estimate for 20% resampled data	0.0388	0.3731	0.4945	0.2780	0.471	0.264	-	0.291

In this study, the error metrics used to quantitatively compare the performance of competing models were mean bias error (MBE), mean absolute error (MAE), root mean square error (RMSE) and mean square error (MSE). Considering the error metrics as given in equations (8, 9, 10 and 11), model 1 had the lowest error term when compared to other models (Table 4). The error metrics of the models whose correlation coefficients were greater than 0.5 (i.e., R^2) in terms of their performance were tested with resampled data set and then compared. Model 1 had the lowest error in terms of the MBE, MAE, RMSE and MSE values when compared to Models 2 and 3. For Models 4 and 5, the predictors (i.e. the explanatory variables) failed to explain the predicted variable (BaP) when tested with 20% resampled and new data because of their weak correlation coefficient of 0.436 and 0.471 respectively. The predictive R-squared (i.e. cross validation: Q^2_{cv}) value for Model 1 (0.791) was also noticed to be outstanding and far above the 0.5 recommended for any quantitative structure-activity relation model [9], [30]. In the same vein, the difference between the R^2 and Q^2_{cv} for a fitted model must be very

minimal (i.e. < 0.3). Thus for model 1, the difference between the R-squared (0.792) statistic and Q^2_{cv} (0.792) statistic was 0.001, which was far less than the 0.1 recommended for an over-fitted model [43]. The difference was also less than 0.3, recommended for a quantitative structure-activity relation (QSAR) model [43]. The Q^2_{cv} value for 20% resampled data for model 1 was calculated to be 0.722. Another internal cross-validation done in this study was the residual plot shown in Fig.3, which showed that the error distribution around the centre-line for model 1 was fairly random (i.e. homoscedastic). While models 2, 3, 4 and 5, displayed an unequal error distribution (i.e. heteroscedastic) with clear patterns, which violates the assumption of a linear model.

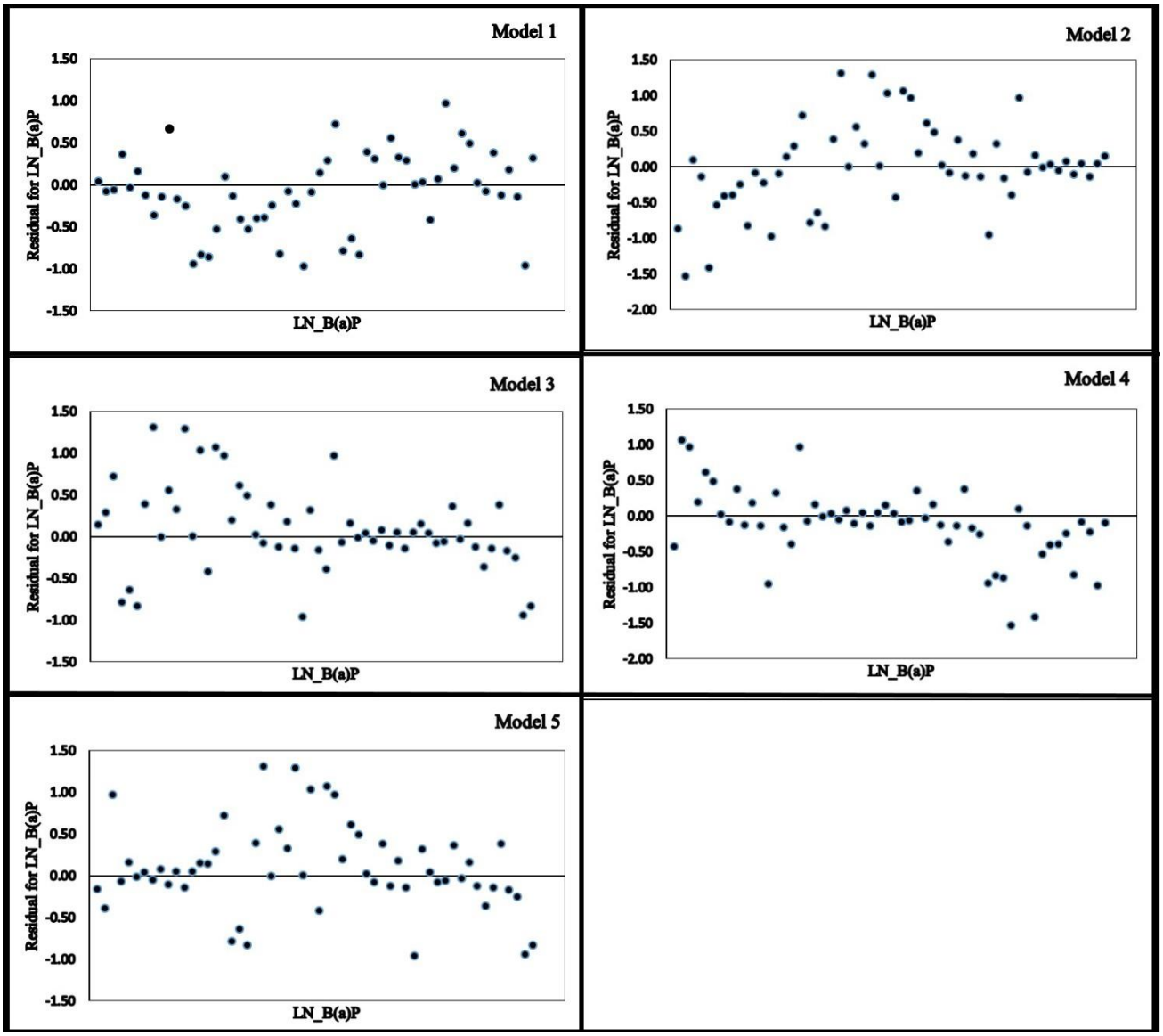

Fig.3. The scattered plot of the residuals for the five models.

3.0.2.2 External validation

The error estimate calculated by using different mathematical metrics and the correlation coefficient(s) for models 1, 2, 3, 4 and 5 respectively are shown in Table 3. The external cross validation (Q^2_{ext}) as reported by many researchers, Q^2_{ext} for a good model when tested with new data must not be less than 0.5 (i.e. 50% predictability power) [27], [22], [6]. Also, the difference between the cross-validation value of the fitted model and new data must be minimal. Therefore from the results of the trained models, model 1 had a Q^2_{ext} value of 0.601 when tested with 20% of the data (new). This revealed that the model was able to predict 60% of new data as against the 79% calculated from the fitted model. Thus, suggesting that the model had a good performance. Considering model 2, the Q^2_{ext} value of 0.860 (86%) for the 20% resampled data was found to be higher than 65% calculated for the fitted model, which suggests an over estimated value arising from an over fitted model. Models 2, 3, 4 and 5 were not consistent in predicting 20% of the resampled and new data. Evaluating the error metrics (MBE, MAE, RMSE and MSE) of individual models with respect to their performance when tested with an independent data (i.e. new data set), models 1 and 3 were quite outstanding as they showed a minimum comparable error values when compared to their fitted model. However model 1 was chosen as the best for this study because of its higher adjusted R- squared (0.746) and predictive R-squared (Q^2_{cv}) value of 0.791. The adjusted R-squared value (0.746) for model 1 revealed that the explanatory variables (i.e. the predictor) used in this study was able to explain 74.6% of the response variable (i.e. the predicted). The predicted R-squared statistic (i.e. internal cross validation: Q_{cv}^2) value of 0.791 in this study showed that model 1 has the ability to predict new set of data with an accuracy of not more than 79.1%.

Table 5 Model hypotheses tests.

Source	Model 1*		Model 2		Model 3		Model 4		Model 5	
	Sig.	Observed Power[b]	Sig.	Observed Power[b]	Sig.	Observed Power[b]	Sig.	Observed Power[b]	Sig.	Observed Power[b]
Corrected Model	0.0000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
Intercept	0.0001	0.995	0.000	0.990	0.000	1.000	0.000	0.992	0.005	0.829
lnT	0.0002	0.986	0.000	0.982	0.000	0.999	0.000	0.985	0.009	0.776
lnWS	0.0000	1.000	0.006	0.821	0.000	0.993	0.000	0.997	0.000	0.984
$lnPM_{2.5}$	0.0003	0.978	0.000	0.971	0.000	0.998	0.000	0.984	0.022	0.649
lnUVR	0.0931	0.390	0.707	0.066	0.577	0.085	0.740	0.062	0.725	0.064
lnT x $lnPM_{2.5}$	0.0004	0.967	0.000	0.965	0.000	0.997	0.000	0.978	0.027	0.616
lnP xlnWS x$lnPM_{2.5}$	0.0078	0.786	0.044	0.530	0.020	0.662	0.006	0.820	0.231	0.220
lnT xlnWS x $lnPM_{2.5}$	0.0199	0.662	0.080	0.419	0.049	0.508	0.017	0.688	0.380	0.139
Total mean variable contribution	0.0174	0.824	0.120	0.679	0.0923	0.749	0.109	0.788	0.199	0.492

Dependent Variable: lnB(a)P

[a] Computed using alpha = .01

[b] Computed using alpha = .05

Model 1: R Squared = .791 (Adjusted R Squared = .746)*

Model 2: R Squared = .647 (Adjusted R Squared = .570)

Model 3: R Squared = .734 (Adjusted R Squared = .675)

Model 4: R Squared = .673 (Adjusted R Squared = .602)

Model 5: R Squared = .630 (Adjusted R Squared = .549)

3.0.3 Model application to NE, SE and SW data

Once the selected model was internally and externally validated, it was necessary to apply the log-linear model (LLM) to the remaining data collected in the same period of study from other zones to further validate the model performance. Out of the 108 samples from the NE, SE and SW data, 22 points were screened out because of outliers using the Z-score value of ±3. Most of the rejected points calculated for B(a)P concentration were far less than that predicted by the model. This may be a result of wet deposition of the $PM_{2.5}$ bound PAHs in the wet season or errors arising during sample(s) extraction. Thus the final set was 86 samples.

The plots of predicted B(a)P values against the experimental B(a)P values for training and test set are presented in Fig. 4.

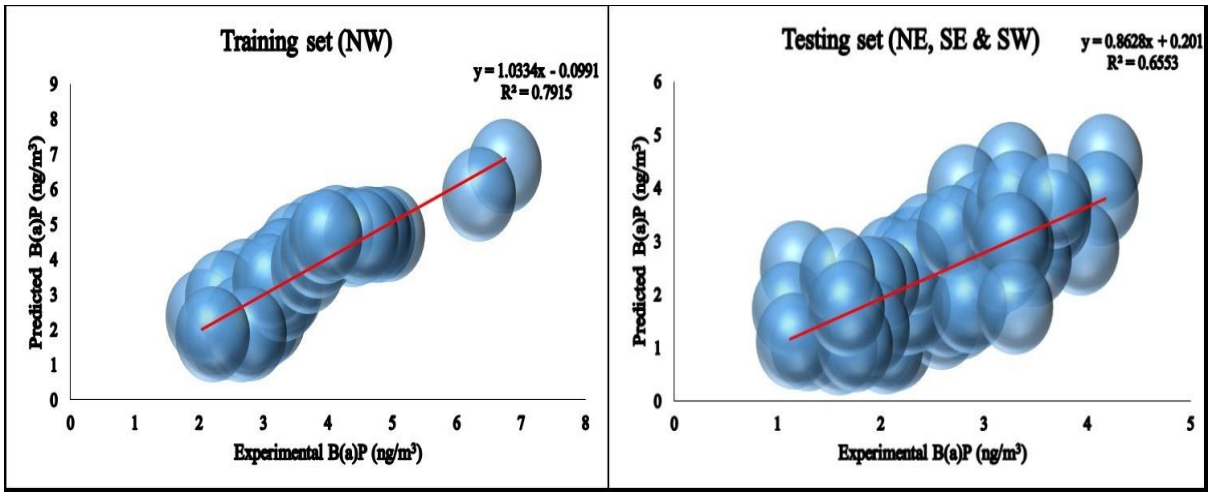

Fig. 4. The plot of Predicted B(a)P Values against the Experimental B(a)P values for Training and Test Sets

A linear relationship was observed in the plot between the experimental and predicted values of the training set (R^2 = 0.792). The fact that all these results were in agreement with a model validation benchmark presented in Table 6, is a confirmation of the reliability, robustness, and stability of the developed model according to Tropsha and his research group as table [42], [11].

Table 6. Benchmark for Model Validation

Validation tools	Interpretation	Acceptable model Value	selected model value	Remarks
	Co-efficient of determination	≥0.6	0.792	Pass
R^2adj	Adjusted R-squared	>0.6	0.745	
P (95%)	Confidence interval at 95% confidence level	<0.05	0.017	Pass
Q^2_{cv}	Cross-Validation Co-efficient (training set)	>0.5	0.791	Pass
No of the samples (training set)	Minimum number of samples for training a GLM model	≥40	72(NW data)	Pass
R^2-Q^2_{cv}	Difference between R^2 and Q^2_{cv} (internal validation)	≤0.3	0.001	Pass
$RMSE_{cv}$	Root mean square error of cross-validation (training set)	NR	0.484	-
MAE_{cv}	Mean absolute error cross validation (training set)	NR	0.400	-
MBE_{cv}	Mean bias error cross-validation (training set)	NR	0.000	-
No of External test-set	Minimum number of external test sets	≥5	108(NE. SE & SW data)	Pass
R^2Test-set (Q^2_{ext})	Co-efficient of determination of external test set (external validation)	≥0.5	0.722	Pass
$RMSE_p$	Root mean square error of prediction (external validation)	NR	0.657	-
MAE_p	Mean absolute error of prediction (external validation)	NR	0.655	-
MBE_p	Mean bias error of prediction (external validation)	NR	-0.152	-

NR: Not recommended

Olasupo and his research also reported that the slope (K) of the regression line between the experimental values and the predicted values in an external validation should be in the range of 0.85 to 1.15 (i.e. $0.85 \leq k \leq 1.15>$) [30]. Thus in this study, the slope (0.8628) between the predicted and the experimental values of the test-set: NE, SE and SW (external validation) were in agreement with that reported by Edacha and his group for a robust model [12].

The line graph shown in Fig. 5, revealed the relationship between experimental and predicted B(a)P for both training and test set. It was evident from the training set (NW-data) that the experimental values (independent variable) in the fitted model were explained by predicted values (dependent variable) to a reasonable extent because of the interaction and closeness between them. In the test set the similarity in trend between the experimental and predicted data (i.e. data generated by the selected model) further validates the model performance.

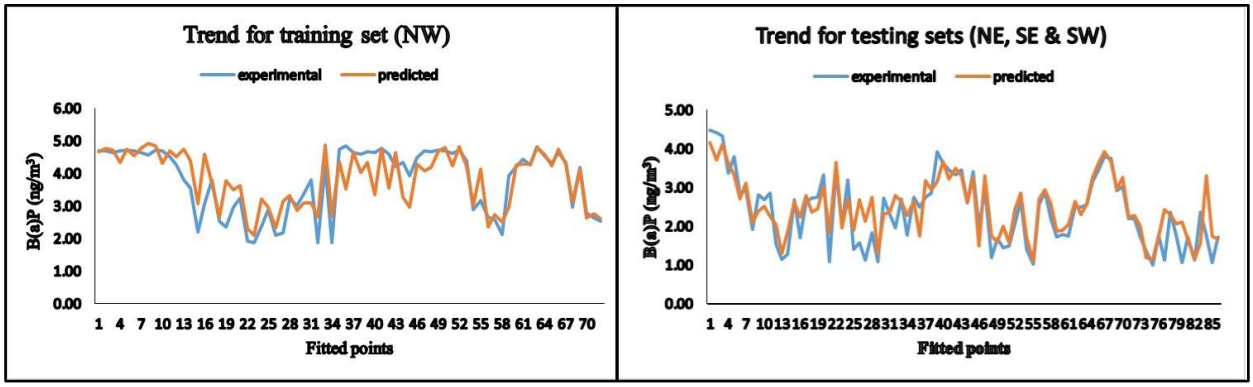

Fig. 5. Line Graph of the Experimental and Predicted Bap for both Training and Test Set.

Thus the model can be said to be robust and can predict B(a)P with reasonable accuracy of 60-72% (Tables 4 and 5) in a new location where similar anthropogenic activities thrived.

4.0 Conclusion

This research aimed at building the first predictive model for the prediction of B(a)P in Nigeria Cities using Benin City as a case study. The internal and external validation of the model coupled with the residual analysis confirmed that the log-linear model (i.e. a special type of the generalized linear model-GLiM) is a promising method for the use of meteorological and particulates ($PM_{2.5}$) data to predict an organic compound such as B(a)P. The prediction ability of the chosen model was confirmed by the R-squared ($R^2 = 0.6553$) obtained when it was applied to the NE, SE, and SW zone data. Thus accounting for 65% predictability of the applied model expected. It can therefore be concluded that the application of the log-linear predictive model offers great potential when exploring the atmospheric concentration of PAHs in the environment where similar activities are taking place.

Author Declarations:

1. **Author Contributions:** James Majebi Okuo designed the research and coordinated it from performing experiments through field sampling to reviewing the manuscript. The first draft of the manuscript was written by Gregory Esosa Onaiwu, coupled with the collection of the field data, performed experiments, modeling and validation, as well as providing input on the interpretation of the results. The manuscript was revised by James Majebi Okuo and Gregory Esosa Onaiwu.
2. **Conflicts of Interest**: We solely declare no conflict of interest.
3. **Funding:** No funding was received for this study except the loan collected from the staff multipurpose co-operative society of Benson Idahosa University and the University of Benin teaching hospital.
4. **Ethics Declaration statement:** Not applicable
5. **Consent to Participate**: Not applicable
6. **Consent for publication**: Not applicable
7. **Data availability statement**: The summary of the data that support the findings of this study is available within the articles. Any other data will be made available from the corresponding author upon reasonable request.

Reference

1. Adeniji AO, Okoh OO, Okoh AI (2018) Analytical Methods for Polycyclic Aromatic Hydrocarbons and their Global Trend of Distribution in Water and Sediment: A Review. Recent Insights Pet Sci Eng. doi: 10.5772/intechopen.71163
2. Aiyesanmi A (2011) Understanding Leaching Behaviour of Landfill Leachate in Benin-City, Edo State, Nigeria through Dumpsite Monitoring. Br J Environ Clim Chang 1:190–200. doi: 10.9734/bjecc/2011/652
3. Araújo IPS, Costa DB, de Moraes RJB (2014) Identification and characterization of particulate matter concentrations at construction jobsites. Sustain 6:7666–7688. doi: 10.3390/su6117666
4. Ashley EK, Ph D, Connor PFO (2017) NIOSH Manual of Analytical Methods (NMAM), 5th Edition Forew o rd. NIOSH Man Anal Methods 1–860
5. Aziakpono OM, Ukpebor EE (2013) Baseline , spatial and temporal variation of total suspended particulate (TSP) matter in Isoko land. 5:129–135. doi: 10.5897/JECE2013.0278
6. Bahmani A, Saaidpour S, Rostami A (2017) A Simple , Robust and Efficient Computational Method for n-Octanol / Water Partition Coefficients of Substituted Aromatic Drugs. 1–14. doi: 10.1038/s41598-017-05964-z
7. C. Moldoveanu S, David V (2019) Derivatization Methods in GC and GC/MS. Gas Chromatogr - Deriv Sample Prep Appl 1–33. doi: 10.5772/intechopen.81954
8. Callén MS, Iturmendi A, López JM (2014) Source apportionment of atmospheric PM2.5-bound polycyclic aromatic hydrocarbons by a PMF receptor model. Assessment of potential risk for human health. Environ Pollut 195:167–177. doi: 10.1016/j.envpol.2014.08.025
9. Callén MS, López JM, Mastral AM (2010) Seasonal variation of benzo(a)pyrene in the

Spanish airborne PM10. Multivariate linear regression model applied to estimate BaP concentrations. J Hazard Mater 180:648–655. doi: 10.1016/j.jhazmat.2010.04.085

10. Cheung K, Daher N, Kam W, Shafer MM, Ning Z, Schauer JJ, Sioutas C (2011) Spatial and temporal variation of chemical composition and mass closure of ambient coarse particulate matter (PM10-2.5) in the Los Angeles area. Atmos Environ 45:2651–2662. doi: 10.1016/j.atmosenv.2011.02.066
11. Edache EI (2015) Multivariate QSAR Study of Indole□-Diketo Acid, Diketo Acid and Carboxamide Derivatives as Potent Anti-HIV Agents. Int J Innov Res Dev 4:374–390. doi: 10.13140/RG.2.2.21955.96801
12. Edache EI, Uzairu A, Mamza PA, Shallangwa GA (2020) A comparative QSAR analysis, 3D-QSAR, molecular docking and molecular design of iminoguanidine-based inhibitors of HemO: A rational approach to antibacterial drug design. J Drugs Pharm Sci 4:21–36. doi: 10.31248/jdps2020.036
13. Ekong F, Michael G, Michael U (2012) Assessing the Effects of Mechanic Activities on Uyo Air Environment. Ethiop J Environ Stud Manag 5. doi: 10.4314/ejesm.v5i1.9
14. Ezeh GC, Obioh IB, Asubiojo OI, Abiye OE (2012) PIXE characterization of PM10 and PM2.5 particulates sizes collected in Ikoyi Lagos, Nigeria. Toxicol Environ Chem 94:884–894. doi: 10.1080/02772248.2012.674133
15. Gnanarajan S (2018) Solutions for Series of Exponential Equations in Terms of Lambert-W Function and Fundamental Constants. J Appl Math Phys 06:725–736. doi: 10.4236/jamp.2018.64065
16. Grönholm T, Annila A (2007) Natural distribution. Math Biosci 210:659–667. doi: 10.1016/j.mbs.2007.07.004
17. Hudda MT, Fewtrell MS, Haroun D, Lum S, Williams JE, Wells JCK, Riley RD, Owen CG, Cook DG, Rudnicka AR, Whincup PH, Nightingale CM (2019) Development and

validation of a prediction model for fat mass in children and adolescents: Meta-analysis using individual participant data. BMJ 366:1–10. doi: 10.1136/bmj.l4293

18. Hussain K, Hoque RR, Balachandran S (2020) Handbook of Environmental Materials Management

19. Juda-Rezler K, Reizer M, Oudinet JP (2011) Determination and analysis of PM10 source apportionment during episodes of air pollution in Central Eastern European urban areas: The case of wintertime 2006. Atmos Environ 45:6557–6566. doi: 10.1016/j.atmosenv.2011.08.020

20. Kim KH, Kabir E, Kabir S (2015) A review on the human health impact of airborne particulate matter. Environ Int 74:136–143. doi: 10.1016/j.envint.2014.10.005

21. Lee DH (2019) Minimizing the Dangers of Air Pollution Using Alternative Facts: A Science Museum Case Study. World Med Heal Policy 11:379–394. doi: 10.1002/wmh3.319

22. Lei T, Li Y, Song Y, Li D, Sun H, Hou T (2016) ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling. J Cheminform 8:1–19. doi: 10.1186/s13321-016-0117-7

23. Liu X, Li C, Tu H, Wu Y, Ying C, Huang Q, Wu S, Xie Q, Yuan Z, Lu Y (2016) Analysis of the effect of meteorological factors on PM2.5-Associated PAHs during autumn-winter in urban Nanchang. Aerosol Air Qual Res 16:3222–3229. doi: 10.4209/aaqr.2016.08.0351

24. Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E (2020) Environmental and Health Impacts of Air Pollution: A Review. Front Public Heal 8:1–13. doi: 10.3389/fpubh.2020.00014

25. Misra S, Osogba O, Powers M (2020) Unsupervised outlier detection techniques for well logs and geophysical data. Elsevier Inc.

26. Obioh IB, Ezeh GC, Abiye OE, Alpha A, Ojo EO, Awolowo O (2013) Toxicological & Environmental Chemistry Atmospheric particulate matter in Nigerian megacities. 37–41. doi: 10.1080/02772248.2013.790970

27. Ojha PK, Mitra I, Das RN, Roy K (2011) Further exploring rm2 metrics for validation of QSPR models. Chemom Intell Lab Syst 107:194–205. doi: 10.1016/j.chemolab.2011.03.011

28. Okuo J, Chiedu I, Anegbe B, Oyibo F, Ojo W (2017) Elemental Characterization and Source Identification of Fine Particulate Matter (PM2.5) in an Industrial Area of Lagos State, Nigeria. Phys Sci Int J 16:1–11. doi: 10.9734/psij/2017/36683

29. Okuo J, Okolo P (2012) Levels of As, Pb, Cd and Fe in Suspended Particulate Matter (SPM) in Ambient Air of Artisan Workshops in Benin City, Nigeria. Bayero J Pure Appl Sci 4:97–99. doi: 10.4314/bajopas.v4i2.19

30. Olasupo SB, Uzairu A, Shallangwa G, Uba S (2019) QSAR analysis and molecular docking simulation of norepinephrine transporter (NET) inhibitors as anti-psychotic therapeutic agents. Heliyon 5. doi: 10.1016/j.heliyon.2019.e02640

31. Orogade SA, Owoade KO, Hopke PK, Adie DB, Ismail A, Okuofu CA (2016) Source apportionment of fine and coarse particulate matter in industrial areas of Kaduna Northern Nigeria. Aerosol Air Qual Res 16:1179–1190. doi: 10.4209/aaqr.2015.11.0636

32. Owoade OK, Fawole OG, Olise FS, Ogundele LT, Olaniyi HB, Almeida MS, Ho MD, Hopke PK (2013) Characterization and source identification of airborne particulate loadings at receptor site-classes of Lagos Mega-City, Nigeria. J Air Waste Manag Assoc 63:1026–1035. doi: 10.1080/10962247.2013.793627

33. Pant P, Lal RM, Guttikunda SK, Russell AG, Nagpure AS, Ramaswami A, Peltier RE (2019) Monitoring particulate matter in India: recent trends and future outlook. Air Qual Atmos Heal 12:45–58. doi: 10.1007/s11869-018-0629-6

34. Del Pero F, Delogu M, Pierini M, Bonaffini D (2015) Life Cycle Assessment of a heavy metro train. J Clean Prod 87:787–799. doi: 10.1016/j.jclepro.2014.10.023

35. Ravindra K, Sokhi R, Van Grieken R (2008) Atmospheric polycyclic aromatic hydrocarbons: Source attribution, emission factors and regulation. Atmos Environ 42:2895–2921. doi: 10.1016/j.atmosenv.2007.12.010

36. Raz R, Roberts AL, Lyall K, Hart JE, Just AC, Laden F, Weisskopf MG (2015) Autism spectrum disorder and particulate matter air pollution before, during, and after pregnancy: A nested case–control analysis within the nurses' health study II cohort. Environ Health Perspect 123:264–270. doi: 10.1289/ehp.1408133

37. Safo-adu G, Ofosu FG, Carboo D, Serfor Y (2014) Health risk assessment of exposure to particulate polycyclic aromatic hydrocarbons at a Tollbooth on a Major Highway . Corresponding Author. Am J Sci Ind Res 5:110–119. doi: 10.5251/ajsir.2014.5.4.110.119

38. Shah ASV, Langrish JP, Nair H, McAllister DA, Hunter AL, Donaldson K, Newby DE, Mills NL (2013) Global association of air pollution and heart failure: A systematic review and meta-analysis. Lancet 382:1039–1048

39. Singh V, Singh S, Biswal A (2020) Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- 19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ' s public news and information

40. Srimuruganandam B, Shiva Nagendra SM (2012) Source characterization of PM 10 and PM 2.5 mass using a chemical mass balance model at urban roadside. Sci Total Environ 433:8–19. doi: 10.1016/j.scitotenv.2012.05.082

41. Taghvaee S, Sowlat MH, Hassanvand MS, Yunesian M, Naddafi K, Sioutas C (2018) Source-specific lung cancer risk assessment of ambient PM2.5-bound polycyclic aromatic hydrocarbons (PAHs) in central Tehran. Environ Int 120:321–332. doi:

10.1016/j.envint.2018.08.003

42. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inform 29:476–488. doi: 10.1002/minf.201000061

43. Wei CC (2016) Comparing single- and two-segment statistical models with a conceptual rainfall-runoff model for river streamflow prediction during typhoons. Environ Model Softw 85:112–128. doi: 10.1016/j.envsoft.2016.08.013

44. Wu SP, Yang BY, Wang XH, Yuan CS, Hong HS (2014) Polycyclic aromatic hydrocarbons in the atmosphere of two subtropical cities in Southeast China: Seasonal variation and gas/particle partitioning. Aerosol Air Qual Res 14:1232–1246. doi: 10.4209/aaqr.2013.01.0015

45. Wu Y, Yang L, Zheng X, Zhang S, Song S, Li J, Hao J (2014) Characterization and source apportionment of particulate PAHs in the roadside environment in Beijing. Sci Total Environ 470–471:76–83. doi: 10.1016/j.scitotenv.2013.09.066

46. Yakubu OH (2018) Particle (Soot) pollution in port harcourt rivers state, nigeria—double air pollution burden? understanding and tackling potential environmental public health impacts. Environ - MDPI 5:1–22. doi: 10.3390/environments5010002

47. Yang TT, Hsu CY, Chen YC, Young LH, Huang CH, Ku CH (2017) Characteristics, sources, and health risks of atmospheric PM2.5-bound polycyclic aromatic hydrocarbons in Hsinchu, Taiwan. Aerosol Air Qual Res 17:563–573. doi: 10.4209/aaqr.2016.06.0283

48. Zhang Y, Zheng H, Zhang L, Zhang Z, Xing X, Qi S (2019) Fine particle-bound polycyclic aromatic hydrocarbons (PAHs) at an urban site of Wuhan, central China: Characteristics, potential sources and cancer risks apportionment. Environ Pollut 246:319–327. doi: 10.1016/j.envpol.2018.11.111