International Journal of Earth, Energy and Environmental Sciences
ISSN: 2517-942X
Vol:12, No:9, 2018

PM10 Prediction and Forecasting Using CART: A Case Study for Pleven, Bulgaria

Snezhana G. Gocheva-Ilieva, Maya P. Stoimenova

Abstract**—**Ambient air pollution with fine particulate matter (PM10) is a systematic permanent problem in many countries around the world. The accumulation of a large number of measurements of both the PM10 concentrations and the accompanying atmospheric factors allow for their statistical modeling to detect dependencies and forecast future pollution. This study applies the classification and regression trees (CART) method for building and analyzing PM10 models. In the empirical study, average daily air data for the city of Pleven, Bulgaria for a period of 5 years are used. Predictors in the models are seven meteorological variables, time variables, as well as lagged PM10 variables and some lagged meteorological variables, delayed by 1 or 2 days with respect to the initial time series, respectively. The degree of influence of the predictors in the models is determined. The selected best CART models are used to forecast future PM10 concentrations for two days ahead after the last date in the modeling procedure and show very accurate results.

Keywords**—**Cross-validation, decision tree, lagged variables, short-term forecasting.

I. Introduction

In many cities around the world, the most harmful pollutants of atmospheric air are particulate matter of size up to 10 microns (PM10). They penetrate into the human body through the respiratory system, with larger particles being retained in the upper respiratory tract, and those with smaller dimensions enter directly into the lungs and damage them. The harmful effect of particulate matter pollution is most pronounced in young children and adults with chronic lung problems, pregnant and neonates [1], [2].

PM10 is the main air pollutant in some eastern European countries, including Poland, Bulgaria, Ukraine and others [3]. In particular, air pollution in Bulgaria is a major environmental problem over the last decade. In order to preserve the cleanness of the atmospheric air and to comply with the requirements of the existing environmental legislation, a National Monitoring System is established by the Executive Environmental Agency [4]. The major sources of the PM10 pollution are mainly domestic heating, especially during winter periods, exhaust gases from gasoline-powered motor vehicles, pollution due to different industrial processes, sanding during the winter periods.

For the control of PM10 concentrations, EU standards and directives apply, according to which the maximum limit values for PM10 in air include an annual average of 40 $\mu g/m^3$ and 24 hour average of 50 $\mu g/m^3$. The daily limit should not be exceeded more than 35 times per calendar year [5], [6].

The air pollution problems are the subject of a large number of scientific studies for modeling and forecasting. Multiple linear and non-linear regression, factor analysis, principal component analysis (PCA), cluster analysis, etc. are widely used multidimensional statistical methods [7]-[10].

Stochastic modeling is another type of preferred approach. With the Box-Jenkins methodology in [11], effective models were developed to study the influence of meteorological factors on PM2.5 particulate matter ultrafine and particulate matter PM10. In [12], factor analysis and ARIMA are combined for modeling average daily concentrations of PM10 over a 10-year period. Also, hybrid ARIMA combined with artificial neural networks, multivariate regression, and other techniques are also used (see [8], [13], [14]). An overview of standard statistical methods and general issues of their application in environmental sciences is addressed in e.g. [15].

Data mining techniques are intelligent data-driven methods to retrieve environmental data and extract patterns and dependences using state-of-the-art high-performance computational algorithms [16]-[18]. Among these methods are: Neural networks, fuzzy logic, support vector machines, CARTs, random forests, etc. Authors of [19] examine the concentration of PM10 in Thessaloniki, Greece over a period of 7 years, depending on meteorological and other variables using multiple linear regression, PCA, neural networks and CART method. Other applications of CART and boosted regression trees to study air pollutants depending on weather conditions, road traffic, and more are presented in [20]. Reference [18] predicts the AQI air quality index in Houston and Los Angeles by modeling data with artificial neural networks, multiple linear regression and vector regression.

In this paper we investigate the capabilities of the highly efficient CART method [21] for modeling and predicting time series of air pollutants. The empirical study is conducted for measured concentrations of PM10 in the city of Pleven, Bulgaria. The particular aims of the study are: 1) obtaining and analyzing mathematical models for the levels of PM10 pollution using CART; 2) exploring and analyzing the impact of meteorological and other factors on pollution; 3) application of models for short-term forecast of PM10.

The modeling was conducted with the Salford Predictive Modeler 8.0 and SPSS software packages.

S. G. Gocheva-Ilieva is with the University of Plovdiv Paisii Hilendarski, 24 Tzar Asen Str., 4000 Plovdiv, Bulgaria (corresponding author, phone: ++359887147238, e-mails: snegocheva@gmail.com, snow@uni-plovdiv.bg).

M. P. Stoimenova is with the University of Plovdiv Paisii Hilendarski, 24 Tzar Asen Str., 4000 Plovdiv, Bulgaria (e-mail: maq.stoimenova@gmail.com).

II. Methodology

A. Study Area and Data

The town of Pleven is located in the central part of the Danubian Plain in the Northwestern Bulgaria (Coordinates: 43°25′N 24°37′E). It is 170 km away from the capital Sofia, 320 km west of the Bulgarian Black Sea Coast and 50 km south of the Danube. Pleven is an administrative center of the Pleven municipality, with about 98 thousand inhabitants. The climate of Pleven is moderately continental, with cold and snowy winters, hot and dry summer months. The average annual temperature is about 13 ° C (55.4 ° F).

B. Data

The study is based on the measured average daily concentrations of air pollutant PM10 in Pleven collected over a period of 6 years, from 1 January 2011 to 31 December 2016. Meteorological time series for minimum and maximum temperatures, wind speed, air pressure, relative air humidity, precipitation, and cloud cover status are also used. The variables and their groups are given in Table I. The first variable PM10 is considered as dependent (response) and other 14 variables are predictors. For further analysis the dependent variable and some predictors are lagged with one or two days back to the current day. There were 10.9% missing data only for PM10, which were replaced by linear interpolation in all analyses. The final data count is N = 2190 cases (observations).

TABLE I
Variables Used for the Construction of CART Models

Variable	Unit	Description
*PM*10	$\mu g/m^3$	PM10 daily average concentration
*PM*10<1>	$\mu g/m^3$	One day lagged *PM*10
*PM*10<2>	$\mu g/m^3$	Two days lagged *PM*10
t		Time, Ordinal
MONTH		Number of month, categorical
min_temp	C°	Minimum daily temperature
max_temp	C°	Maximum daily temperature
wind_speed	m/s	Wind speed
precipitation	%	Precipitation
humidity	%	Relative humidity
pressure	mb	Air pressure
cloud-cover	%	Cloud-cover
min_temp<1>	C°	One day lagged *min_temp*
min_temp<2>	C°	Two days lagged *min_temp*
wind_speed<1>	m/s	One day lagged *wind_speed*

Fig. 1 presents graphically the initial data for average daily PM10 concentrations (upper part) and the respective minimum and maximum daily temperatures (lower part). For PM10, multiple exceedances of the permissible daily average limit of 50 $\mu g/m^3$ (marked with a horizontal line) are observed, mainly during winter periods. Generally, the behavior of the PM10 time series corresponds in an inverse proportional way to the minimum and maximum daily temperatures.

For the six-year period, the measured mean value of PM10 pollution is 49 $\mu g/m^3$. From 2011, this indicator varies as: 52.3, 45.4, 41.7, 51.1, 53.9, 48.4 $\mu g/m^3$, which is systematically exceeding the permissible annual average of 40 $\mu g/m^3$. A maximum daily concentration of PM10 of 363 $\mu g/m^3$ has been reached. In total, there are 705 (over 32%) exceedances of the prescribed limit value for the average daily values for PM10 of 50 $\mu g/m^3$. We can conclude that PM10 is a very problematic pollutant for the town of Pleven.

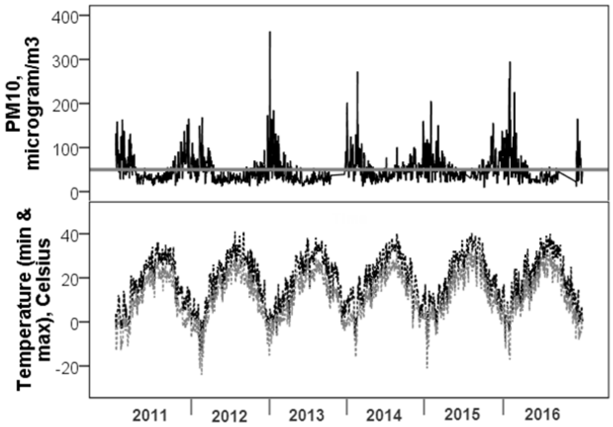

Fig. 1 Daily concentrations of PM10, maximum and minimum daily temperatures in the town of Pleven from 2011 to 2016

C. Short Description of CART Method

CART method, also known as Decision Trees is introduced in [21] for solving classification or regression predictive modeling problems in the class of machine learning data mining methods. The idea of the method is to divide the selected initial learning sample of cases into non-overlapping sub-sets by classifying similar cases. A recursive tree of nodes is grown, for example, binary. At each step of the algorithm, the current node is divided into two child nodes. For pre-defined splitting criteria, the algorithm stops. These criteria could be for example, a minimum number of cases in a parental and in a child node, fixed maximum depth of the tree, reaching a certain accuracy, etc. As a result, the initial set of cases will be distributed to the terminal nodes of the decision tree. The predicted value for each case is equal to the mean value of the dependent variable for the cases classified in its respective terminal node. To obtain an optimal tree, the splitting at each step is performed according to a rule of the type: "*predictor value* $\leq \theta_j$?", where θ_j is some value of this predictor. The model error at each step *K* is equal to the current sum of the squared residuals, e.g.

$$S(K) = \sum_{t=1}^{N} \left(O_t - P_{K,t} \right)^2 , \qquad (1)$$

where O_t is observed value of the dependent variable (PM10 in our case) at time *t* and $P_{K,t}$ is its corresponding predicted value. The optimal CART model gives a minimum of (1). For comparison between models with the same predictors in the CART method, a relative error is used, defined by

$$Rel.Err. = \frac{S(K)}{S(0)}, \tag{2}$$

where the number K of the terminal nodes varies and the denominator is the sum (1) for the entire sample.

Usually, CART models could be validated by the standard machine learning V-fold cross-validation (CV) [22]. In the procedure of V-fold CV, the sample is randomly divided into V equal sub-samples. Each of the sub-samples is used to test the model, and the rest of the data is used as a learning sample, the process being repeated V times.

Main advantages of the CART method are:

- Capable to handle equally well numerical and categorical data and combinations of these
- Independent of the type of distribution
- Able to detect complex non-linear dependencies
- Simple to understand and interpret
- Easy to predict new response values

Some shortcomings of the method are:

- Applicable with at least of 50 observations
- In some cases, does not give enough accurate models, comparing to other methods.

III. Results and Discussion

A. Modeling Settings

To obtain the best models for PM10 concentrations we applied CART algorithm with 10-fold cross-validation and different combinations of predictors, including lagged variables (see Table I). From all obtained models, constructed with a given set of predictors, the models within 1 standard error (1SE) of accuracy are considered [21]. Other settings for pre-defined criteria are minimum number of cases in a parent node 10 and in a child node 5, which are considered as usual [22].

B. Model Selection Criteria

To estimate the accuracy and adequacy of the CART models we use the relative CART error defined in (2), and the standard indicators of the Root Mean Square Error (RMSE) and the coefficient of regression (R) given by the expressions

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(O_t - P_t\right)^2} \tag{3}$$

$$R = \frac{Cov(O,P)}{Std(O).Std(P)} \tag{4}$$

The criteria used for selecting the best model are: the smallest relative CART error and RMSE, and the highest value of R. When model indicators are close, the simpler model is chosen.

C. Model Construction

A number of models were built and analyzed, by varying and excluding some of the predictors from Table I with the least influence on the models. Seven best models have been selected to meet the requirements for the best model. The results are shown in Table II. The first model $M1$ has the smallest relative error of 0.387 but its remaining indicators are of lower quality (relatively bigger RMSE and smaller R). Among the selected seven models with smaller RMSE and higher R are the $M5$ and $M7$ models. Since $M7$ is more complex (with the most terminal nodes), we can choose the simpler model, $M5$. Here, we also have to note that the selected best $M5$ model has an optimized number of the most important predictors (only 8, see Table III).

The predictor sets used and their relative importance in the seven selected models are given in Table III. From this table, it is immediately observed that with the highest importance on the models with 100% is the pollution from the previous day (lagged variable PM10 <1>) and also from the previous two days (PM10 <2>). This influence is of stochastic nature.

The time variable t and the current month also have a stable participation in all models about 16-18% and 43-44%, respectively (except model $M7$). Of the meteorological variables as expected the most significant influence on PM10 pollution shows the minimum daily average temperature and wind speed. The rest of meteorological factors are less influential.

TABLE II
Statistical Indicators of the Selected CART Models

Model	Terminal nodes	R Learn	Rel.Err. Learn	Rel.Err. Test	RMSE
*M*1	74	0.901	0.187	0.387	12.702
*M*2	107	0.909	0.175	0.417	12.261
*M*3	131	0.912	0.168	0.417	12.035
*M*4	116	0.910	0.173	0.418	12.217
*M*5	158	0.915	0.163	0.414	11.853
*M*6	123	0.910	0.171	0.414	12.149
*M*7	218	0.917	0.177	0.420	11.694

TABLE III
Variable Importance for Selected CART Models[1)]

Predictor	*M*1	*M*2	*M*3	*M*4	*M*5	*M*6	*M*7
*PM*10<1>	100	100	100	100	100	100	100
*PM*10<2>	-	45.7	47.1	46.6	47.2	46.8	46.6
t	15.9	15.8	18.0	15.3	19.0	18.0	-
MONTH	43.9	44.1	43.6	44.3	44.0	43.7	14.1
min_temp	-	32.3	31.1	32.2	34.9	34.5	15.3
max_temp	-	10.9	-	-	-	12.1	10.8
wind_speed	42.6	19.2	20.1	19.7	20.2	19.6	19.3
precipitation	10.6	9.1	10.0	9.7	10.6	9.9	8.4
humidity	-	10.8	14.1	11.3	16.5	14.8	10.7
pressure	-	13.6	14.6	14.0	-	-	14.0
cloud-cover	6.4	6.2	-	6.5	-	-	-
min_temp<1>	48.2	-	-	-	-	-	47.7
min_temp<2>	53.1	-	-	-	-	-	-
wind_speed<1>	-	-	-	-	-	-	26.6

All values are percentages of maximum importance, which is assumed to be 100%. Missing values denote excluded predictors in the analyses.

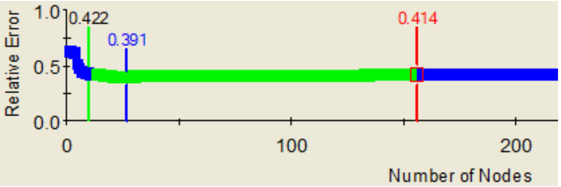

Fig. 2 Relative error curve for models built with *M5* predictor set from Table III

Fig. 2 shows the plot of the relative error curve of the models built using *M5* predictor set (see variables in the respective column of Table III). In green color (middle part) are marked all models with relative errors within 1SE. The optimal model has a relative error of 0.391 and 50 terminal nodes, and the maximum model (*M5*) has 158 terminal nodes and *R*=0.915.

The regression tree topology of the best selected model *M5* is shown in Fig. 3. The largest predicted value is indicated by the arrow. Its value is 244 μg/m^3. The series of rules that classify the cases in this terminal node are as follows:

$$(PM10<1> \; > \; 61.30) \;\&\&\; (wind\_speed \le 2.45) \;\&\&$$
$$(PM10<1> \; > \; 165.26) \;\&\&\; (PM10<2> \le 118.68),$$

where && denotes conjuncture.

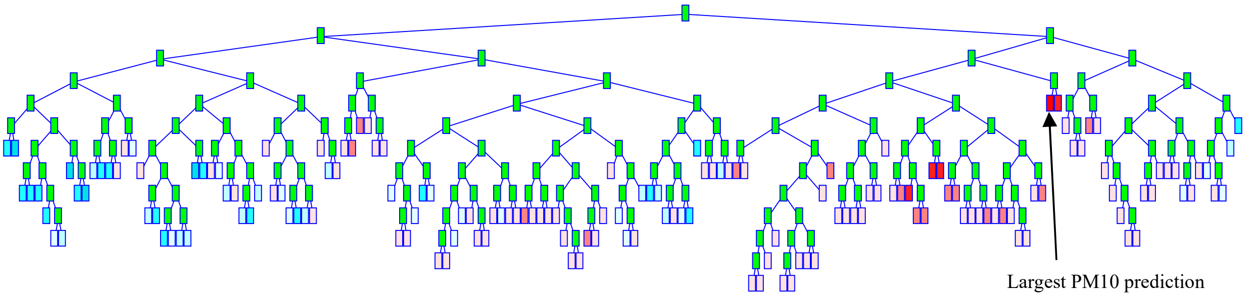

Fig. 3 Regression tree topology of model *M5*

D.Model Diagnosis and Evaluation

All obtained models were checked for consistency by analyzing their model errors. For time series, usually the autocorrelation function (ACF) of the residuals could be examined. For model *M5* this is illustrated in Fig. 4. It is observed that ACF coefficients of residuals are small enough in the due confidence limits.

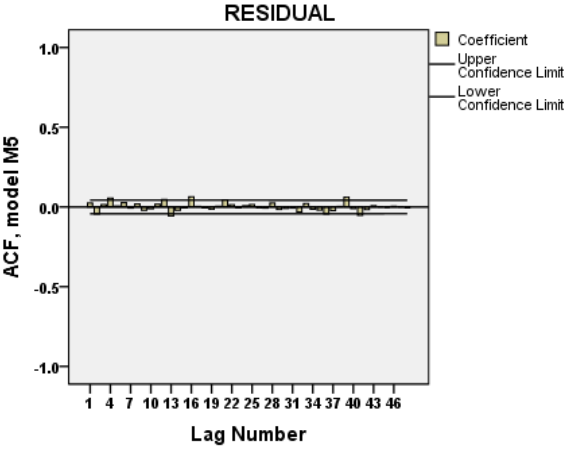

Fig. 4 ACF of residuals for model *M5*

An intuitive way to assess model quality is to graphically compare the observed with the predicted values ones. Fig. 5 presents the predictions obtained by the selected best CART model *M5* compared with the measured PM10 values. The goodness-of-fit measure R^2 (coefficient of determination) is equal to 0.837. It can be assumed that the model describes about 84% of the data.

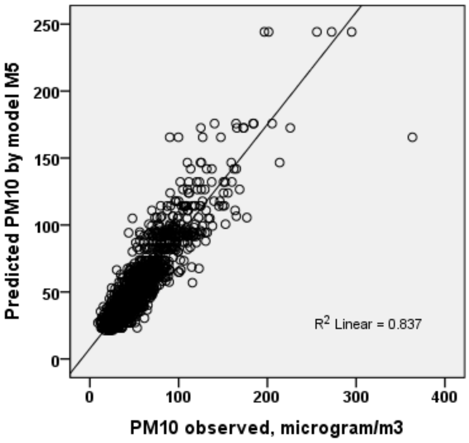

Fig. 5 Comparison of the measured values of PM10 with their predictions from model *M5*

For practical application of the models they can be evaluated depending on the number of matches with respect to the prescribed upper daily limit for PM10 pollution of 50 μg/m^3. Table IV presents the number of correctly and incorrectly classified cases by model *M5*.

In our case the observed PM10 values above the limit are 704 out of 2190. The selected model *M5* correctly classified

566 or 80.4% exceedances over 50 μg/m^3, as well as of 1412 or 95% of the data below the threshold. The total number of correctly predicted PM10 values represents 90.3%. Improperly predicted below the average daily limit are 138 and over 50 μg/m^3 are 74. From these results it is found that the selected model *M*5 demonstrates very good accuracy in the prediction of the measured PM10 values with respect to the threshold value.

TABLE IV
CONTINGENCY TABLE FOR SELECTED CART MODEL *M*5 AND PM10

		Predicted			
		<50 μg/m^3	>=50 μg/m^3	Total	%Obs
Obs	<50	**1412**	74	1486	95.0%
	>=50	138	**566**	704	80.4%
	Total	1550	640	2190	**90.3%**
	%Predicted	91.1%	88.4%		

E. Application of the Model for Two Days Forecasting

We also checked the quality of the models by comparing their forecasts with known PM10 values for two days ahead in the time series (for January 1 and 2, 2017) that are not involved in the construction of the models. The comparison results for models *M*2, *M*3, and *M*4 are shown in Fig. 6, and for models *M*5, *M6*, and *M*7 are shown in Fig. 7, respectively.

The first five values in Figs. 6 and 7 are for the last days in the initial data sample used in modeling procedure - from December 27 to December 31, 2016, and the last two days values are the real PM10, compared to the models predictions and forecasts for January 1 and 2, 2017. These values are separated by a vertical line. Horizontal lines indicate requirements from the air quality standard with a threshold value of 50 μg/m^3.

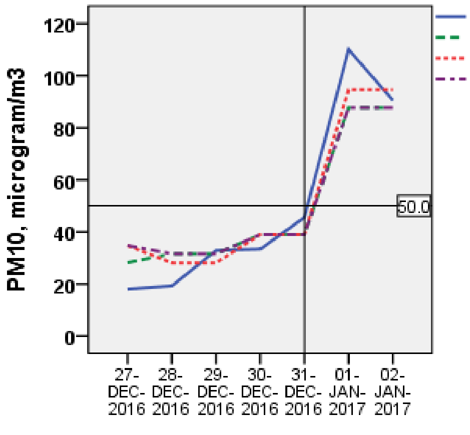

Fig. 6 Comparison of the measured values of PM10 with the predictions and forecasts of models *M*2, *M*3, and *M*4

It is observed that all models give very good predictive and forecasted results. This also applies to the correct prognosis of lower and higher values compared to the threshold value. It can be inferred that the CART approach is very suitable for properly classifying pollution and can be successfully applied to alert the population. The resulting tree is easily interpreted and applied for forecasting for a short period of time in the future for which fairly accurate weather forecasts are known.

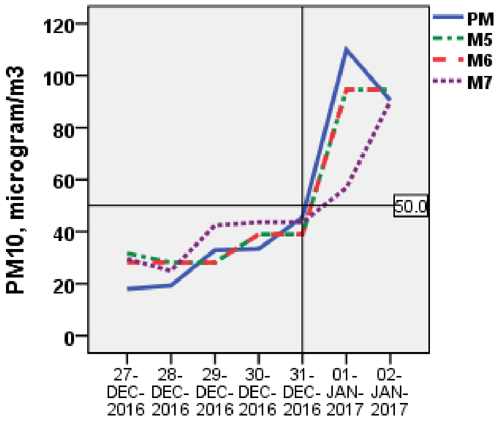

Fig. 7 Comparison of the measured values of PM10 with the predictions and forecasts of models *M*5, *M*6, and *M*7

ACKNOWLEDGMENT

This study was partially supported by the Scientific Fund of the University of Plovdiv Paisii Hilendarski, Bulgaria, project MU17-FMI-003.

REFERENCES

[1] *Health Effects of Particulate Matter. Policy Implications for Countries in Eastern Europe, Caucasus and Central Asia*. World Health Organization, 2013. <www.euro.who.int/__data/assets/pdf_file/0006/189051/Health-effects-of-particulate-matter-final-Eng.pdf>

[2] A. Seaton, D. Godden, W. MacNee, and K. Donaldson, "Particulate air pollution and acute health effects," *The Lancet*, vol. 345, no. 8943, pp. 176-178, 1995.

[3] *Air quality in Europe - 2014 report*. European Environment Agency, Publications, 19 Nov 2014. <http://www.eea.europa.eu/publications/air-quality-in-europe-2014/at_download/file>

[4] Executive Environment Agency, Bulgaria. <http://eea.government.bg/en>

[5] *Air Quality Standards*: *Environment*. European Commission. <http://ec.europa.eu/environment/air/quality/standards.htm>

[6] "Directive 2008/50/EC of the European Parliament and of the council of 21 May 2008 on ambient air quality and cleaner air for Europe," *Official Journal of the European Union*, L 152/1, 2008.

[7] S. Abdullah, M. Ismail, and S. Y. Fong, "Multiple linear regression (MLR) models for long term PM10 concentration forecasting during different monsoon seasons," *Journal of Sustainability Science and Management*, vol. 12, no. 1, pp. 60-69, 2017.

[8] A. Vlachogianni, P. Kassomenos, A. Karppinen, S. Karakitsios, and J. Kukkonen, "Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki," *Science of The Total Environment*, vol. 409, no. 8, pp. 1559-1571, 2011.

[9] K. Y. Ng and N. Awang, "Multiple linear regression and regression with time series error models in forecasting PM10 concentrations in Peninsular Malaysia," *Environ Monit Assess*, vol. 190, no. 63, pp. 1-11, 2018. https://doi.org/10.1007/s10661-017-6419-z

[10] I. Zheleva, E. Veleva, and M. Filipova, "Analysis and modeling of daily air pollutants in the city of Ruse, Bulgaria," in *AIP Conference Proceedings,* vol. 1895, 030007, 2017.

[11] L. Jian, Y. Zhao, Y. P. Zhu, M. B. Zhang, and D. Bertolatti, "An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China," *Science of The Total Environment*, vol. 426, pp. 336-345, 2012.

[12] P. W. G. Liu, "Simulation of the daily average PM10 concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis," *Atmospheric Environment*, vol. 43. pp. 2104-2113, 2009.

[13] M. Zickus, A. J. Greig, and M. Niranjan, "Comparison of four machine learning methods for predicting PM10 concentrations in Helsinki, Finland," *Water, Air, & Soil Pollution: Focus*, vol. 2, pp. 717-729, 2002.

[14] S. G. Gocheva-Ilieva, A. V. Ivanov, D. S. Voynikova, and D. T. Boyadzhiev, "Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach", *Stochastic Environ Res Risk Assess*, vol. 28, no. 4, 1045-1060, 2014.

[15] D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, 3nd ed. Amsterdam: Elsevier, 2011.

[16] F. Biancofiore, M. Busilacchio, M. Verdecchia, B. Tomassetti, E. Aruffo, S. Bianco, S. Di Tommaso, C. Colangeli, G. Rosatelli, and P. Di Carlo, "Recursive neural network model for analysis and forecast of PM10 and PM2.5," *Atmos Poll Research*, vol. 8, no. 4, pp. 652-659, 2017.

[17] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Amsterdam: Morgan Kaufmann, Elsevier, 2016.

[18] S. S. Ganesh, P. Arulmozhivarman, and R. Tatavarti, "Forecasting air quality index using an ensemble of artificial neural networks and regression models," *Journal of Intelligent Systems*, 2017. https://doi.org/10.1515/jisys-2017-0277

[19] T. Slini, A. Kaprara, K. Karatzas, and N. Moussiopoulos, "PM10 forecasting for Thessaloniki, Greece," *Environ Modell Softw*, vol. 24, no. 1, pp. 559-565, 2006.

[20] W. Choi, S. E. Paulson, J. Casmassi, and A. M. Winer, "Evaluating meteorological comparability in air quality studies: Classification and regression trees for primary pollutants in California's South Coast Air Basin," *Atmospheric Environment*, vol. 64, pp. 150-159, 2013.

[21] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont: Wadsworth, 1984.

[22] A. J. Izenman, *Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning*, New York: Springer, 2008.