ECOLOGIA BALKANICA

2020, Vol. 12, Issue 1 June 2020 pp. 171-177

Modeling and Forecasting of Air Pollution with Particulate Matter PM2.5 Depending on Weather Conditions in Urban Areas - A Case Study from Plovdiv, Bulgaria

Maya P. Stoimenova-Minova[*]

University of Plovdiv „Paisii Hilendarski", Faculty of Mathematics and Informatics, Department of Applied Mathematics and Modeling, 24 Tzar Assen Str., 4000 Plovdiv, BULGARIA
[*]Corresponding author: maq.stoimenova@gmail.com

Abstract. One of the main global environmental issues with direct impact on public health is preserving clean air in urban areas. The latest report by the European Environment Agency details the progress made towards meeting the air quality standards comparing it against the official requirements set out in EU directives. The objective of this scientific study is to model air pollution in the city of Plovdiv with fine particulate matter smaller than 2,5 micrograms per cubic meter (PM2.5). A CART method is used and based on best results obtained forecasts are made for future particulate matter pollution for 2 days ahead. Data from actual measurements taken between 1 February 2014 and 30 June 2019 and data from meteorological measurements are used. This study is an alternative to the official reports by the Regional Inspectorate of Environment and Water for the city of Plovdiv, which enables forecasting future pollution and its prevention.

Key words: Particulate matter PM2.5, air pollution, pollution forecast, Classification and regression trees (CART) method.

Introduction

Air pollution is a serious issue relevant to many European countries. According to the latest report by the European Environment Agency (EEA, 2019), progress has been made in achieving legal requirements on air quality. The most harmful air pollutants are fine particulate matter up to 10 microns in diameter (PM10) and up to 2,5 microns (PM2.5). The two types of pollutants cause severe respiratory diseases, lung damage, allergies, etc. Smaller particulate matter PM2.5 make their way directly to lung alveoli and cause diseases such as cancer, heart disease, etc. The main source of harmful particles are the burning of solid fuels by households and large factories, gases emitted by motor vehicles, especially diesel engines, etc. It has been proven through epidemiological and toxicological studies that particulate matter are harmful to human health. For this reason, it is necessary to monitor concentration levels and to take the necessary measures to meet EU requirements and to eliminate any violations.

There are numerous scientific studies about the impact of harmful emission on human health, as well as studies on mortality related to these problematic pollutants (Herman et al., 2020; Maji et al., 2017; Cox et al., 2013).

The environmental issue of air pollution is subject to studies around the world. The goals of the researchers include finding pollution sources, namely (Saraga et al., 2019; Ehsanzadeh et al., 2016). The Box-Jenkins method has been used to build stochastic models (Jian et al., 2012), which consider the influence of meteorological factors on PM2.5 and PM10 emissions. Various types of methods are used to study air quality through modeling using artificial neural networks, CART, fuzzy sets and others (Ivanov & Gocheva-Ilieva, 2013; Prakash et al., 2011).

In the past decade, urban air pollution in Bulgaria has been a serious problem. Constant control and monitoring of harmful emissions is performed by the National Environment Agency. Plovdiv and Sofia are among the cities with relatively high levels of PM10 pollution in Europe in recent years (WHO, 2015). In Bulgaria, there is relatively little scientific literature to analyze the condition of atmospheric air in urban environments (Gocheva-Ilieva & Ivanov, 2019; Gocheva-Ilieva et al., 2019; Veleva & Zheleva, 2018).

The main objective of this paper is to study the dependence of PM2.5 concentration on meteorological conditions by building CART models to be used as the basis of short-term air pollution forecasts. The proposed approach can be an independent corrective to aid local governments and communities in warning about exceeded permissible limits for fine particulate matter.

Material and Methods

Study area. The city of Plovdiv is the second largest city in Bulgaria after the capital Sofia. Its population is around 350,000 people. It is situated along the two banks of the Maritza river. The climate in the city is transitional-continental with an average annual temperature of 12,3 °C. The average annual relative air humidity is 73%, which is highest in December. During the cold months there are often fogs. Plovdiv is a city with weak winds and low altitude.

Data and initial statistical processing. The analysis performed on the state of air in the city of Plovdiv is based on actual measurements by an automated measuring station “Kamenitza” located in the city center. The station is located in a predominantly residential area with moderate vehicle traffic. The data used for the scientific study are average daily measurements of the PM2.5 pollutant over a period of about 5 and a half years – from 1 February 2014 to 30 June 2019. This paper also utilizes measurement data for 8 meteorological indicators; these are: minimum average daily temperature (minT, °C), maximum average daily temperature (maxT, °C), wind speed (wind_speed, km/h), wind direction (wind_direction, rad(WDI)), precipitation (precip, %), air humidity (humidity, %), atmospheric pressure (pressure, mb) and cloud cover (cloud_cover, %). In order to account for the periodic nature of the variable wind_direction it is transformed using the formula for WDI:

$$WDI = 1 + \sin\left(wind\_direction + \pi / (k-1) \right),$$

where *k* is the number of different wind directions and in this case *k*=16.

The autoregression variables of PM2.5 , minT, maxT and wind_speed are used to build the models. They are denoted respectively by PM2.5 <1> - these are the fine particulate matter concentrations measured on the previous day, minT<1>, minT<2> - values of the minimum temperature measured on the previous day and two days before, maxT<1> - maximum average daily temperature on the previous day and wind_speed<1> - wind speed on the previous day. Since the dependent variable is a time series, variables taking into account time are also dded – month and month_day.

CART Method. CART (Classification And Regression Trees) method was developed in 1984 in a monograph (Breiman et al., 1984). The method is actively used in almost all scientific fields for the classification and investigation of dependencies. With a quantitative type variable y and independent predictors $X = (X_1, X_2, ..., X_p)$ the CART algorithm builds a binary tree structure for the observations by splitting the multivariate case space into non-intersecting regions. Starting from the root of the tree, which contains all cases, at each step, the cases are split into two upon satisfying a preset rule of the type $x_k \leq \theta_{k,j}$. If the rule is met, the cases are classified in a left child node of the current node, and the rest – into the right child node. The process of growing the tree is ceased according to criteria specified by the researcher, for example minimum number of cases in a parent node (m_1) and in a child node (m_2), depth of the tree, etc (Izenman, 2008; Steinberg & Colla, 1995).

The model $\hat{y}$ can be written down as:

$$\hat{y}(\mathbf{X}) = \sum_{l=1}^{m} \hat{y}(\tau_l) I_{[X \in \tau_l]}, \ \hat{y}(\tau_l) = \overline{y}_l, \ X \in \tau_l,$$

where $\tau_l, l = 1, 2, ..., m$ are terminal nodes of the tree, m – their number, $I_{[X \in \tau_l]}$ is the function, which traces the route from the root to the terminal node, $\overline{y}_l$ is the mean value of cases, classified in the terminal node τ_l, which is also the predicted value.

The quality of the models is assessed using the Root Mean Square Error (RMSE) and the coefficient of determination according to formulas:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}, \quad R^2 = \frac{\sum_{t=1}^{N}(\hat{y}_t - \overline{y})^2}{\sum_{t=1}^{N}(y_t - \overline{y})^2},$$

$$t = \overline{1, N},$$

where N is the number of cases, $\hat{y}_t$ is the predicted value at any given moment t, $\overline{y}$ is the mean value of y.

When selecting the best model, the guiding principle is to achieve the least root mean square error and the highest value for the coefficient of determination R^2, i.e. the degree of approximation of the model to actual data.

Modeling is performed using Salford Predictive Modeler suite (SPM, 2016) and IBM SPSS (IBM Corp, 2013).

Results

Table 1 presents the results of the initial processing of the data. The total number of observations for the pollutant PM2.5 is N=1948, and the missing data are 27, which represents about 1,37%. There are no missing values in the data about meteorological measurements.

Table 1. Descriptive statistics of the initial data on PM2.5 concentrations in the city of Plovdiv.

Statistics	PM2.5 $\mu g/m^3$	minT (°C)	maxT (°C)	wind_ speed (*km/h*)	WDI	precip (%)	humi dity (%)	pres sure (mb)	cloud_ cover (%)
N	1948	1975	1975	1975	1975	1975	1975	1975	1975
N miss	27	-	-	-	-	-	-	-	-
Mean	25.53	6.69	17.07	6.46	2.88	1.04	0.677	1017	0.30
Minimum	2.3	-17	-10	2	0.9	0	0.31	991	0
Maximum	253	22	40	26	11.6	47.6	0.99	1039	1

During the examined period, a maximum of 253 $\mu g/m^3$ was reached for PM2.5, while the legal requirements do not permit exceedance of more than 25 $\mu g/m^3$ during the year. (EC, 2008; 2015). The permitted values for PM2.5 concentrations in the air are effective as of 1 January 2015 with a maximum average annual limit of 25 $\mu g / m^3$, without any exceedance throughout the year. As of 1 January 2020, average annual limits of 20 $\mu g / m^3$ are in force. For the years between 2014 and 2019, the average values are respectively 28,79; 29,35; 27,44; 27,32; 19,17; 17,55 $\mu g / m^3$. The analysis indicates a systemic exceedance of the permissible limit.

Fig. 1 presents a graph of the observed average annual data over the examined period. The horizontal line in the graph corresponds to the average permissible limit of 25 $\mu g / m^3$. Numerous exceedances of the specified legal requirement are observed. The main reasons for the high levels of harmful emissions given in the regional reports are the use of solid fuels for heating during the winter period and the poor quality of fuels.

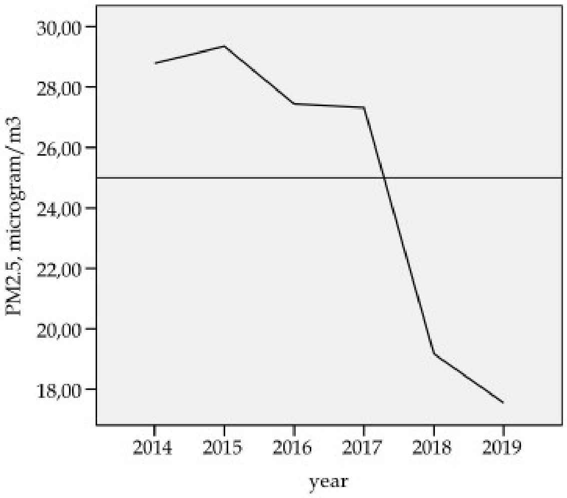

Fig. 1. Graph of measured average yearly data on PM2.5 concentrations in the city of Plovdiv. The horizontal line indicates the permissible limit of 25 $\mu g / m^3$

Construction of CART models and their analysis. When building CART models, the main goal is to find the dependence of the high levels of the PM2.5 air pollutant on meteorological conditions. Limitations are set on the minimum number of cases in the parent node (m_1) and the child node (m_2). Following numerous preliminary analyses, the values of 10 are preset for m_1 and 5 for m_2. The obtained models are denoted by $M(m_1, m_2)$.

Out of a large number of models built during the study, 3 optimal models are selected that match the requirements for best fit. Table 2 presents their basic characteristics, which indicate the extent to which the selected models approximate the actual data, the number of end nodes, and the errors reported during their build. According to the above conditions for best model, Table 2 shows the conclusion that M3 provides the best results. It describes over 82% of actual data and has the lowest RMSE=10,519.

Table 2. Summary of the obtained optimal CART models for PM2.5.

Varia ble	Model	(m1, m2)	Number of Terminal Nodes	R2 Lea rn	RMSE
PM2.5	M1	(10.5)	34	0.794	11.297
PM2.5	M2	(10.5)	47	0.814	10.727
PM2.5	**M3**	(10.5)	51	**0.821**	**10.519**

Table 3 shows the percentage ratio of meteorological variables, included in the build of the three optimal models selected. The influence of the most important indicator assumes a weight of 100%. Some of the meteorological conditions are excluded when building the models since during the analyses it was found that these have no influence. The excluded meteorological data are precipitation quantity, atmospheric pressure and cloud cover. It is obvious that for all models, the measured PM2.5 concentrations on the previous day have the greatest weight and the next most stable factor in terms of influence is the minimum average daily air temperature two days before. It can be concluded that the selected models demonstrate stability in their structure. For the best model, it was found that the greatest influence (100%) on harmful emissions is that of PM2.5 concentrations measured on the previous day, followed by the measured minimum temperature two days before, with the third being minimum temperature on the previous day, etc.

Fig. 2 shows the general structure of the obtained regression tree using model M3, which includes 51 end nodes as per the CART tree build rules. Terminal node 51 classifies the cases with the highest values of PM2.5. The rules to reach this node starting from the tree root are as follows: PM2.5<1> >48,78 $\mu g / m^3$; PM2.5<1> >128,14 $\mu g / m^3$; PM2.5<1> >184,31 $\mu g / m^3$. The predicted value of each terminal node is the arithmetic mean of cases in it.

Table 3. Variable importance of the predictors for obtained models.

Predictors	**Scores**		
	M1	**M2**	**M3**
PM2.5<1>	100	100	100
minT<2>	19.96	20.54	20.85
minT<1>	19.44	20.22	20.27
humidity	12.91	12.85	13.09
weather	9.90	11.09	10.70
wind_speed<1>	9.80	9.74	9.91
minT	5.55	6.51	6.62
month	5.25	-	5.34
maxT	3.18	-	-
wind_speed	2.95	-	-
month_day	-	7.93	-
maxT<1>	-	5.63	5.33
wind_speed<2>	-	-	6.84

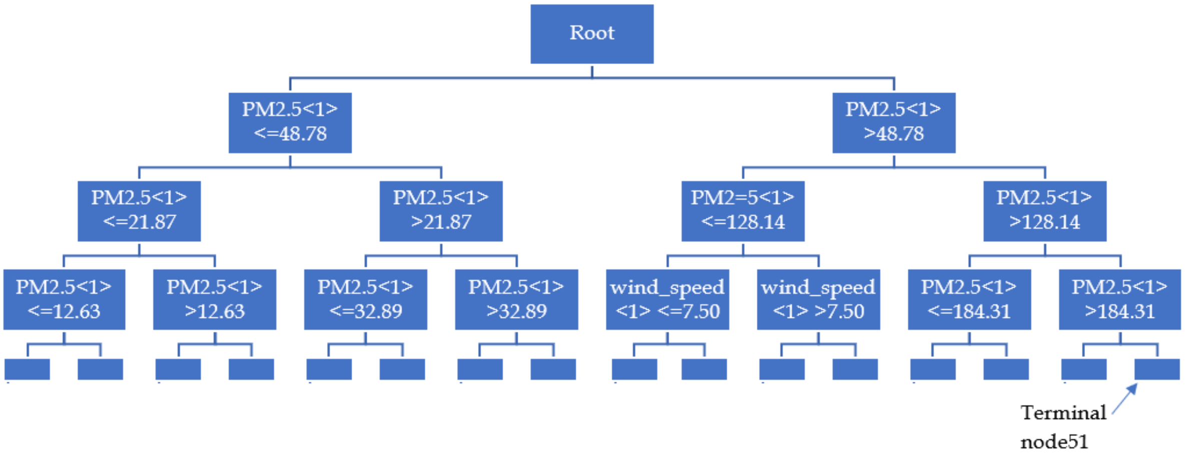

Fig. 2. Upper part of topology of the binary regression CART tree of the model M3.

Application of the models for forecasting future concentrations. Fig 3 shows a graphic of the predicted PM2.5 concentrations for 2 days ahead (1 and 2 July 2019), using data models up to 30 June 2019. The forecasts for the two days are made using actual measured data, which are not included when building the CART model but these are compared against the values predicted by the selected model. The graphic shows that the selected model approximates actual measured values of the air pollutant.

Discussion

The study uses actual measurements as data for air pollution with PM2.5 in the city of Plovdiv. The data are for the period from 1 February 2014 to 30 June 2019. The analysis shows that during the studied period there are many exceedances of the limits for healthy air quality. Modeling and forecasting the atmospheric pollutant are performed using the classification and regression trees (CART) method. Three optimal CART models of the level of fine particulate matter smaller than 2,5 $\mu g / m^3$ depending on meteorological conditions are built and analyzed. The graphic forecast of the model compared against actual measured PM2.5 concentration show very good approximation. It is found that the selected model displays very good qualities for short-term forecasts of particulate matter air pollution for 2 days ahead.

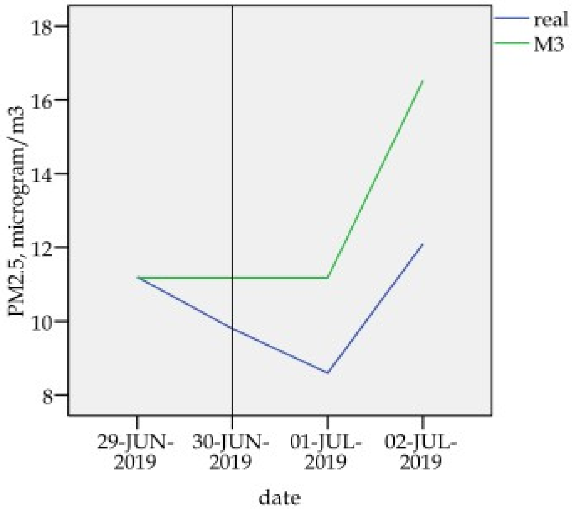

Fig. 3. Comparison of measured and predicted PM2.5.

According to the reports by the Regional Inspectorate of Environment and Water in Plovdiv on air quality (RIOSV Plovdiv, 2019), significant influence on the high levels of particulate matter in the city is due to temperature inversions, large number of days without any wind and the fogs. These factors cause accumulation and longer duration of suspension of pollutants in the air. Vehicle traffic emissions also have a significant negative impact on air quality.

Conclusions

This study shows the results of a statistical analysis of air quality in the city of Plovdiv. The obtained results show that legal limits that guarantee public health are exceeded. The proposed approach demonstrates the influence of meteorological conditions on air pollution. The selected method is suitable for forecasting these in order to prevent and control harmful emissions in the air within urban areas.

The presented study shows that the selected approach is suitable for predicting future pollution and its prevention.

Acknowledgements

The author wish to acknowledge the financial support of the National Program "Young Scientists and Postdoctoral Students" of the Ministry of Education and Science in Bulgaria, 2020.

References

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth Intern, Belmont.

Cox, T., Popken, D., & Ricci, P. (2013). Associated with Short-Term Acute Daily Mortality Rates: Results from One Hundred United States Cities. *International Dose-Response Society, 11*(3), 319-343. doi: 10.2203/dose-response.12-034.Cox.

EC. (2008). 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. 2008. *Official Journal of the European Union*, *L152,* 1-44. Retrieved from eur-lex.europa.eu

EC. (2015). *Air Quality Standards.* European Commission. Environment. Retrieved from ec.europa.eu

EEA. (2019). *Air quality in Europe - 2019 report*. European Environment Agency. Publications. Retrieved from eea.europa.eu

Ehsanzadeh, A., Nejadkoorki, F., & Khodadoostan., S. (2016). A study on the most important factors affecting the concentration of particulate matter smaller than 10 microns (PM10) using principal component regression. *Journal of Research in Environmental Health, 2*(2), 154-164. doi: 10.22038/jreh.2016.7584.

Gocheva-Ilieva, S., & Ivanov, A. (2019). Assaying stochastic SARIMA and generalized regularized regression for particulate matter PM10 modeling and forecasting. *International Journal of Environment and Pollution, 66*, 41-62. doi: 10.1504/IJEP.2019.104520.

Gocheva-Ilieva, S. G., Voynikova, D. S., Stoimenova, M. P., Ivanov, A. V., & Iliev, I. P. (2019). Regression trees modeling of time series for air pollution analysis and forecasting. *Neural Computing and Applications, 31*(12), 9023-9039, doi: 10.1007/s00521-019-04432-1.

Herman, D., Wingen, L., Johnson, R., Keebaugh, A., Renusch, S., Hasen, I., Ting, A., & Kleinman, M. (2020). Seasonal effects of ambient PM2.5 on the cardiovascular system of hyperlipidemic mice. *Journal of the Air and Waste Management Association, 70*(3), 307-323. doi: 10.1080/10962247.2020.1717674.

IBM Corp. (2013). *SPSS IBM Statistics.* Vers. 22. Retrieved from ibm.com

Ivanov, A. V., & Gocheva-Ilieva, S. G. (2013). Short-time particulate matter PM10 forecasts using predictive modeling techniques. *Fifth Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences. Melville. American Institute of Physics, AIPCP*, *1561*, 209-218. doi: 10.1063/1.4827230.

Izenman, A. (2008). *Modern Multivariate Statistical Techniques Regression, Classification and Manifold Learning.* New York.

Jian, L., Zhao, Y., Zhu, Y. P., Zhang, M. B., & Bertolatti, D. (2012). An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. *Science Total Environmental, 426*, 336–345. doi: 10.1016/j.scitotenv.2012.03.025.

Maji, J., Dikshit, A., & Deshpande, A. (2017). Disability - adjusted life years and economic cost assessment of the health effets related to PM2.5 and PM10 pollution in Mumbai and Delhi, in India from 1991 to 2015. *Environmental Science and Pollution Research, 24*, 4709-4730. doi: 10.1007/s11356-016-8164-1.

Prakash, A., Kumar, U., Kumar, K., & Jain, V. (2011). A wavelet-based neural network model to predict ambient air pollutants' concentration. *Environmental Modeling and Assessment., 16*(5), 503-517. doi: 10.1007/s10666-011-9270-6.

RIOSV Plovdiv. (2019). *Report on the state of air quality.* Retrieved from plovdiv.riosv.com

SPM. (2016). *Salford Systems Data Mining and Predictive Analytics Software Modeler*, Vers. 8.0. Retrieved from salford-systems.com.

Saraga, D., Tolis, E., Maggos, T., Vasilakos, C., & Bartzis, J. (2019). PM2.5 source apportionment for the port city of Thessaloniki, Greece. *Science of The Total Environment, 650*(2), 2337-2354. doi: 10.1016/j.scitotenv.2018.09.250.

Steinberg, D., & Colla, P. (1995). *CART: Tree-Structured Non-Parametric Data Analysis.* San Diego.

Veleva, E., & Zheleva, I. (2018). CARH models for particulate matter PM10 air pollutant in the city of Ruse, Bulgaria. *10th Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences. Melville. American Institute of Physics, AIPCP,* 2025(040016). doi: 10.1063/1.5064900.

WHO. (2015). *Air pollution*. World Health Organization. Retrieved from who.int.

Received: 10.04.2020
Accepted: 08.05.2020