Regression trees modeling and forecasting of PM10 air pollution in urban areas

Cite as: AIP Conference Proceedings **1895**, 030005 (2017); https://doi.org/10.1063/1.5007364
Published Online: 12 October 2017

M. Stoimenova, D. Voynikova, A. Ivanov, S. Gocheva-Ilieva, and I. Iliev

Regression Trees Modeling and Forecasting of PM10 Air Pollution in Urban Areas

M.Stoimenova[1,a)], D. Voynikova[1,b)], A. Ivanov[1,c)], S. Gocheva-Ilieva[1,d)], and I. Iliev[2,e)]

[1]*Paisii Hilendarski University of Plovdiv, 24 Tzar Asen str., 4000Plovdiv, Bulgaria*
[2]*TechnicalUniversity of Sofia, branch Plovdiv, 25 Tsanko Dyustabanov str., 4000 Plovdiv, Bulgaria*

[a)]Corresponding author: maq.stoimenova@gmail.com
[b)]desi_sl2000@yahoo.com
[c)]aivanov_99@yahoo.com
[d)]snegocheva@gmail.com
[e)]iliev55@abv.bg

Abstract. Fine particulate matter (PM10) air pollution is a serious problem affecting the health of the population in many Bulgarian cities. As an example, the object of this study is the pollution with PM10 of the town of Pleven, Northern Bulgaria. The measured concentrations of this air pollutant for this city consistently exceeded the permissible limits set by European and national legislation. Based on data for the last 6 years (2011-2016), the analysis shows that this applies both to the daily limit of 50 micrograms per cubic meter and the allowable number of daily concentration exceedances to 35 per year. Also, the average annual concentration of PM10 exceeded the prescribed norm of no more than 40 micrograms per cubic meter. The aim of this work is to build high performance mathematical models for effective prediction and forecasting the level of PM10 pollution. The study was conducted with the powerful flexible data mining technique Classification and Regression Trees (CART). The values of PM10 were fitted with respect to meteorological data such as maximum and minimum air temperature, relative humidity, wind speed and direction and others, as well as with time and autoregressive variables. As a result the obtained CART models demonstrate high predictive ability and fit the actual data with up to 80%. The best models were applied for forecasting the level pollution for 3 to 7 days ahead. An interpretation of the modeling results is presented.

INTRODUCTION

Preserving the air quality in urban areas is one of the main global environmental problems which have a direct impact on public health. In many countries around the world, one of the most harmful air pollutants is fine particulate matter up to 10 microns in diameter (PM10). According to a large number of scientific studies and official sources, increased concentrations of PM10 cause severe respiratory diseases, damaging lung tissue, allergies, *etc.*(*e.g.,* [1, 2] and the literature cited therein). The harmful effect of particulate matter pollution is worst for young children and adults with lung disease. In Europe as a whole, this pollutant is problematic mostly in some East European countries among which Bulgaria, Poland, Ukraine, *etc*. Official information was recently published in the latest European Union report for 2014 [3]. More specifically, urban air pollution in Bulgaria has been a significant ecological problem in the last decade. In order to improve the air quality and to meet the requirements of the applicable environmental legislation, a National System for Environmental Monitoring is responsible, subsystem "air" [4]. But as the official reports of many municipalities such as Sofia, Plovdiv, Ruse, *etc.* indicate, there is a trend for persisting harmful levels of PM10 despite various measures taken to reduce these.

Application of Mathematics in Technical and Natural Sciences
AIP Conf. Proc. 1895, 030005-1–030005-10; https://doi.org/10.1063/1.5007364
Published by AIP Publishing. 978-0-7354-1579-9/$30.00

The considered air pollution issues are subject to a large number of scientific investigations by teams from different fields. Various methods are applied for modeling and examining the available data, highlighting different aspects of the pollution processes in separate cities, urban areas and larger regions. Among the most important objectives of the modeling is the forecasting of pollution concentrations in regions where air quality is a problem. Stochastic models using the Box–Jenkins method with transfer functions are constructed in [5] in order to investigate the influence of meteorological factors on ultra-fine particulate matter PM2.5 and PM10 particulate matter concentrations. In [6] factor analysis and ARIMA with transfer functions are used to model the mean daily concentrations of PM10 over a period of 10 years. There are numerous publications investigating air quality by applying artificial neural network methods, multilayer perception, wavelet analysis, classification and regression trees (CART), support vector machines (SVMs), fuzzy sets, and others (see, for example [7-9] and the literature cited therein). In the literature review of this problem, we have to note [10], where the concentration of PM10 in Thessaloníki, Greece is investigated over a period of 7 years in relation to meteorological and other time series. In this paper different modeling methods are applied and compared, such as multivariate linear regression, principle component analysis, neural networks and CART. Statistical models based on multiple regression analysis and CART are used in [11] to model and forecast the levels of ozone and PM10 in Lisbon in Portugal.

In this paper we examine the application of powerful and flexible data mining CART technique for constructing models for analyzing and forecasting the PM10 pollution levels, measured over the period of 6 years, from January 2011 to December 2016 in the town of Pleven, a typical Bulgarian city. The aim of the study is to determine the influence of meteorological data on the mean daily PM10 concentrations, taking also into account three time variables. An additional objective is to investigate in which extend the transformation of the *PM*10 variable into a new variable with almost normal distribution influence the model performance. The application of the models for short-term forecasting of future pollution for one and two days ahead is also studied.

Modeling is performed using Salford Predictive Modeler suite [12] and IBM SPSS [13].

STUDY AREA, DATA AND DATA PRE-PROCESSESING

The city of Pleven is located at an altitude of 150m in the central Danube plain in Northwestern Bulgaria. It is situated 174 km northeast of the capital – Sofia. The city is the administrative center of Pleven Province and has about 98 thousand inhabitants. The climate in Pleven is temperate continental with cold and snowy winters, and hot and dry summers. The average annual temperature is around 13°C (55.4°F).

Pollution monitoring, prevention, and control in the European Union is governed by Directive 2008/50/EC on ambient air quality and cleaner air for Europe and standards for various pollutants [14, 15]. For particulate matter PM10 the limit values for the concentrations are: 24-hour average of $50\,\mu g/m^3$, allowing no more than 35 exceedances within a calendar year, and 40 $\mu g/m^3$ annually – without any exceedance. The prescribed limits by the World Health Organization (WHO) for PM10 are: up to 20 $\mu g/m^3$ annually and up to 50 $\mu g/m^3$ per day [3].

The analysis within this study is based on the measured average daily concentrations of the PM10 air pollutant in Pleven, Bulgaria, over a period of 6 years, from 1 January 2011 to 31 December 2016. The values measured for 8 meteorological variables are also used for the modeling and forecasting: *min_temp* (^{o}C)– minimum average daily air temperature, *max_temp* (^{o}C)– maximum daily temperature, *wind_speed* (m/s) – average wind speed for the day, *wind_direction* (radians) – wind direction, *precipitation* (%), *humidity* (%) – air humidity, *pressure* (mb) – atmospheric pressure, *cloud_cover* (%). In order to account for the periodic nature of the variable wind direction, it is transformed into WDI using the expression

$$WDI = 1 + \sin\left(wind\_direction + \pi/4\right). \quad (1)$$

The descriptive statistics of the main characteristics for all initial data and the additionally transformed variable *tr_PM*10 are given in Table 1. The missing data for *PM*10 are 10.9%, which are replaced using linear interpolation in the analyses.

Table 1 shows that over the six-year period, the average value of *PM*10 is 49 $\mu g/m^3$. By years, this indicator has the following average values since 2011 – 52.3, 45.4, 41.7, 51.1, 53.9, 48.4 $\mu g/m^3$. This is a systemic exceedance of the permissible limit of 40 $\mu g/m^3$ per calendar year. The maximum daily concentration for *PM*10 of 363 $\mu g/m^3$ is reached. In the measured values, there are a total of 705 (over 32%) exceedances over the prescribed threshold limit of 50 $\mu g/m^3$ for the daily mean values for PM10.

TABLE 1. Descriptive statistics of initial and transformed data

Statistics	*PM*10 μg/m^3	*tr_PM*10 μg/m^3	*min_ temp* ºC	*max_ temp* ºC	*wind_ speed* m/s	*WDI*	*precipit ation* %	*humid ity* %	*pressu re* mb	*cloud_ cover* %
N	1951	2190	2190	2190	2190	2190	2190	2190	2190	2190
N missing	239	0	0	0	0	0	0	0	0	0
Mean	48.97	1.93	11.7	19.14	2.73	1.14	1.90	0.67	1017	0.30
Median	40.12	1.93	12.0	20.0	2.2	1.00	0.10	0.67	1017	0.22
Std. Dev.	30.60	0.10	10.31	11.06	1.53	0.69	4.34	0.16	7.37	0.26
Skewness	2.70	0.04	-0.26	-0.27	1.52	-0.13	3.99	-0.12	0.19	1.04
Std.Err. Skewness	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Kurtosis	13.37	0.01	-0.65	-0.82	3.00	-1.48	20.34	-0.67	0.36	0.18
Std.Err. Kurtosis	0.11	0.11	0.11	0.11	0.11	0.10	0.11	0.11	0.11	0.11
Minimum	8.96	1.5	-24	-13	0.40	0	0	0.24	990	0
Maximum	363.36	2.26	32	41	11.6	2	43	0.99	1041	1.00

Figure 1 shows the sequence plots of the part of data variables. The upper section of the figure shows the measured average daily concentrations of PM10, and the lower – the respective maximum and minimum daily temperatures. For PM10 it is obvious that there are multiple exceedances over the permissible average daily limit of $50\,\mu g/m^3$ (indicated by the horizontal line), mostly during the winter. As a whole, these peaks coincide with the lowest values of the minimum and maximum daily temperatures during winter months. Pollution exceedances during summer months are far fewer, mostly during 2014 and 2015. The observed seasonal differences are explained by the different sources of pollution. According to the dispersion modeling of PM10 pollution in the regional inspectorate reports [16], conclusions are drawn that the burning of fossil fuels by household during the heating season is the number one contributor to high PM10 concentrations, and vehicle transport is the second leading factor, including the irregular cleaning of the road network. Industrial pollution has decreased significantly, but background levels of PM10 remain high, resulting from secondary dust deposition and specific regional topography, and variation of climatic and meteorological factors.

The calculated statistical indices in Table 1 lead us to the assumption that most variables have not normal distribution. This applies mostly to variables *PM*10, *min_temp*, *max_temp*, *wind_speed*, *precipitation* and *cloud_cover*. Here, the main indicators are the high absolute values of the ratios of Skewness to Standard Error of Skewness and respectively of Kurtosis to Standard Error of Kurtosis as well as the performed Kolmogorov-Smirnov tests of normality.

In our analyses, we also take into account the influence of the non-normal distribution of the dependent variable *PM*10. In order to improve normality, we use the power transformation of Yeo-Johnson [17] in this case of the non-negative values of the variable with empirically determined in our case coefficient $\lambda = -0.4$ from the expression

$$tr\_PM10 = \frac{(PM10+1)^{\lambda}-1}{\lambda}, \quad PM10 \geq 0, \quad \lambda \neq 0, \quad \lambda \in [-2,2]\,. \tag{2}$$

The resulting statistics for *tr_PM*10 are given in the respective column of Table 1. Figure 2 shows the histograms of *PM*10 and the transformed variable *tr_PM*10. It is observed that the achieved distribution after the transformation by (2) is nearly normal.

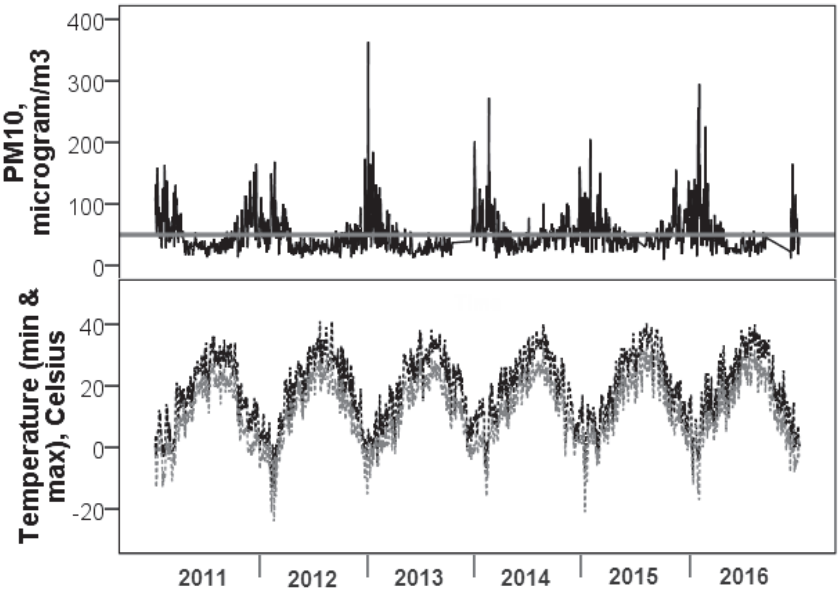

FIGURE 1. Sequence plots of the measured daily concentrations of PM10, maximum and minimum daily temperatures

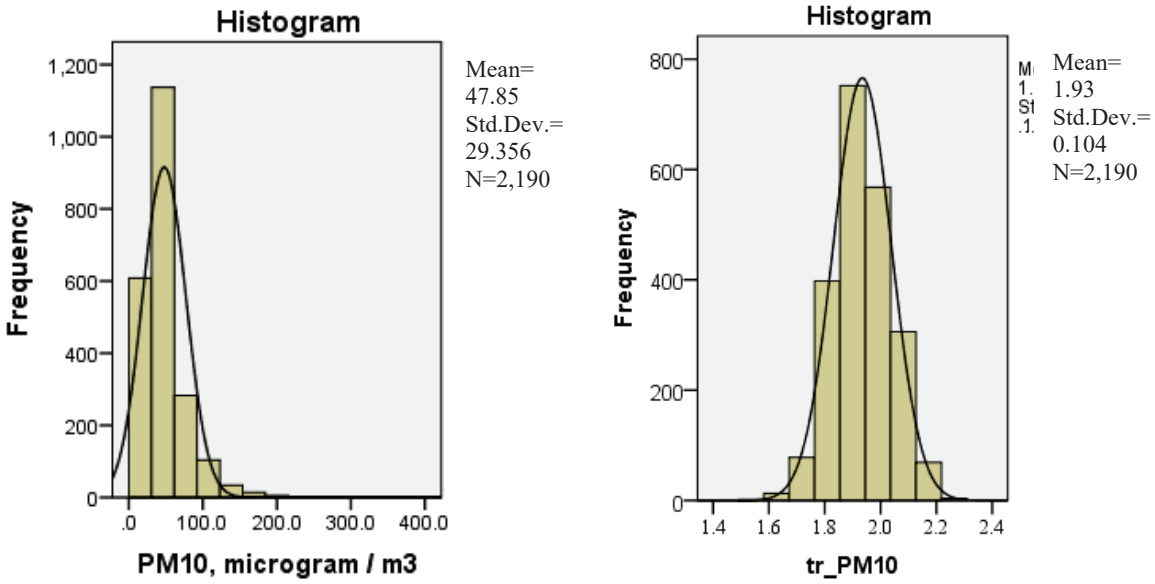

FIGURE 2. Histograms of *PM*10 and the transformed variable *tr_PM*10for Pleven, Bulgaria

BRIEFDESCRIPTIONOFCARTMETHOD

The CART (Classification And Regression Trees) method is proposed in the monograph [18] in 1984. Today, it is used actively for the classification and investigation of relationships in almost all scientific fields, and is considered to be one of the most effective and powerful statistical techniques for data mining in informatics [19]. There are different variations of CART, *e.g.*, CHAID, Exhaustive CHAID, *etc.*

As a regression technique, the CART method is considered as a recursive partitioning regression. The objective is to split the cases (observations) with data into relatively homogeneous (with small standard deviation or minimum total error using the least square measure) terminal nodes and to calculate a mean value of dependent variable at

each terminal node as a predicted value. With a quantitative variable Y and an arbitrary type independent predictors $\mathbf{X}=(X_1,X_2,...,X_p)$, the CART algorithm builds a tree structure for the classification of cases by splitting the multivariate predictor space of cases into non-intersecting sub-regions. By starting from the root of the tree, containing all cases, at each step, the cases are split into two by applying a rule of the type $X_k \le \theta_{k,j}$ for the values of some predictor. If the cases fulfill the rule, they are classified in the left descendant node of the current parent node, and the others – into the right descendant node. The value $\hat{Y}$ predicted by the model at a given node τ is the mean value of Y of the cases classified within the node τ. At each step where two descendant nodes are generated, the selection of a predictor X_k is chosen out of all predictors and all its possible values $\theta_{k,j}$ for the cases within the node so that the total current model leads to the smallest deviance, or smallest relative mean square error. The process of growing the tree ceases according to criteria selected by the researcher, *e.g.*, minimum number of cases in a parent node (m_1) and descendant node (m_2), depth of the tree, *etc.* [20-22] The total regression CART model is compiled by the rules for classification of all values of the independent variable Y in terminal nodes of the tree and their predicted mean values $\overline{Y}_\ell$. The model may be written down as the following equation

$$\hat{Y}(\mathbf{X})=\sum_{\ell=1}^{m}\hat{Y}(\tau_\ell)I_{[\mathrm{X}\in\tau_\ell]},\hat{Y}(\tau_\ell)=\overline{Y}_\ell,\ X\in\tau_\ell \tag{3}$$

where τ_ℓ, $\ell=1,2,...,m$ are terminal nodes, m - their number, $I_{[\mathrm{X}\in\tau_\ell]}$ is the indicator function tracing the route from the root to the terminal node τ_ℓ where $I=1$, if the logic value of the cross-section of the respective rules is true and $I=0$ if it is false. With the help of (3), it is easy to find the forecast value of the new observation, if its respective values of the predictors are known. Following the rules, the terminal node is found, where the new observation is classified and its forecast is the mean value of the dependent variable in the respective node.

BUILDING CART MODELS OF *PM*10 AND *TR_PM*10

The goal is to build regression CART models in order to determine the relationship between the levels of PM10 pollution and the eight meteorological variables. Instead of the variable for wind speed we used the transformed WDI according to the expression (1). Since the dependent variable is a time series, we add three other predictors to these, taking time into account: *year_month* (represented as increasing sequence of real numbers, see for instance [23]). The limitations for a minimum number of cases per parent node (m_1) and descendant node (m_2) following a large number of preliminary analyses are specified in two variants – 20 and 10 for m_1, and 10 and 5 for m_2, respectively. The obtained models of *PM*10 are denoted by $\mathrm{M}(m_1,m_2)$. All models are also built under the same conditions for the transformed variable tr_PM10, denoted hereafter by $\mathrm{tr\_M}(m_1,m_2)$.

When selecting the best model, we are guided by the least relative error of model data from the measurements, according to [21, 22]. What is more, since we seek for a regression type of model, we also take into consideration the models having the largest value of the coefficient of determination R^2. Furthermore, we investigate whether the distribution of the residuals is normal. The adequacy of the models requires very good prediction of the measured actual observations and the forecast future pollution. In order to compare the accuracy of the forecasts, we use known values of PM10 for two days ahead in the time series (1 and 2 January 2017), which have not been used when building the models.

RESULTSANDDISCUSSION

Four optimal models are selected which meet the specified conditions in the previous section. Their main characteristics are given in Table 2. Table 2 shows that the CART model M1 for *PM*10 is relatively simple with 169 terminal nodes but accounts for about of 61% of the actual measured data according to the value of R^2=0.61. The comparison with model M2 gives an advantage to the latter, which has a significantly lower relative error. For the models of *tr_PM*10, a similar performance ratio is valid. These have a higher coefficient of determination R^2 of the respective models of *PM*10 within 6-7% with a very small difference in tree complexity in terms of number of terminal nodes. The relationship is also analogical for the respective relative errors. The overall comparison shows

that the best model of the selected is *tr*_M4, accounting for up to 78% of the measured data with the smallest relative error 0.222.

TABLE 2. Summary of the obtained optimal CART models for *PM*10 and *tr_PM*10

Variable	Model	(m_1, m_2)	Number of the Terminal Nodes	R^2Learn	Relative Error
*PM*10	M1	(20,10)	169	0.6086	0.391
	M2	(10,5)	349	0.7274	0.273
*tr_PM*10	tr_M3	(20,10)	173	0.6773	0.323
	tr_M4	(10,5)	354	0.7776	0.222

The upper part of topological structure of the resulting regression tree with the model *tr*_M4 having 354 terminal nodes is given in Figure 3 with the rules for growing the CART tree. The highest terminal node is the first one at the left side of the figure. Starting from the root of the tree, the classification rules in this node are as follows:*min_temp*<= 7.5, *wind_speed*<= 2.45, wind_speed <= 1.1, *min_temp*<= –12.5. Thus, we obtain that the highest exceedances in measured *PM*10 concentrations are obtained at a relatively low wind speed below 1.1 m/s and a minimum temperature below -12.5 ºC.

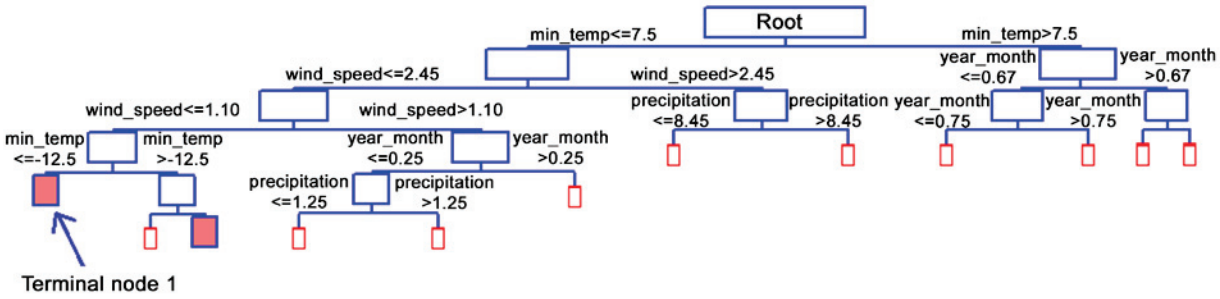

FIGURE 3. Upper part of topology of the binary regression CART tree of the model *tr*_M4 with 354 terminal nodes and 9 predictors for the transformed variable *tr_ PM*10

Table 3 shows the relative importance percentages of the predictors participating in the four models selected. These percentages are relative to the most important one, whose influence is assumed to be 100%.

TABLE 3. Variable importanceof the predictors for obtained models

Predictor	Scores			
	M1	M2	*tr*_M3	*tr*_M4
min_temp	100	100	100	100
wind_speed	99.52	93.24	48.66	50.81
year_month	89.08	83.39	98.97	94.43
max_temp	75.16	73.03	75.03	77.60
Pressure	58.06	61.80	69.65	70.54
Humidity	38.25	43.29	34.59	43.46
Precip	26.55	27.47	32.04	39.72
cloud_cover	26.19	32.75	30.25	36.35
WDI	12.87	14.17	12.91	17.11

It is obvious that the largest weight in all models is that of the minimum average daily air temperature (variable *min_temp*). The next predictor with stable influence is the time variable, accounting for the year and month (*year_month*) whose importance varies between 83 to 94%. Next in terms of stability but having less influence are the maximum average daily air temperature (*max_temp*) and air pressure. An exception is *wind_speed*, which has an influence decreasing from 99.5% in M1 to 51% in tr_M4 as the model number goes up. The remaining predictors have weaker influence. As a whole, it can be concluded that the selected models demonstrate stability in their structure.

Table 4 provides more detailed results from the analysis of the four obtained CART models. For each model, statistics are provided for the two nodes with the highest predicted values of the dependent variable (*PM*10 and *tr_PM*10, respectively). The description of the classification rules for falling into the terminal node is given from the root to the node. This makes it possible to trace the conditions for forecasting pollution.

TABLE 4. Summary of the obtainedCARTmodels*)

Model	Selected terminal node	Mean	Number of cases	Classification rules in the selected terminal node
M1	17	124.94	11	*min_temp*≤ 6.5, wind_speed≤ 2.45, *wind_speed*> 1.1, *year_month*≤ 0.17, *min_temp*> -8.5, *max_temp*> 11.5, *cloud_cover*≤ 0.325
M1	1	133.42	14	*min_temp*≤ 6.5, *wind_speed*≤ 2.45, *wind_speed*≤ 1.1, *min_temp*≤ -6.5
M2	1	170.95	5	*min_temp*≤ 6.5, *wind_speed*≤ 2.45, *wind_speed*≤ 1.1, *min_temp*≤ -12.5
M2	2	163.66	5	*min_temp*≤ 6.5, *wind_speed*≤ 2.45, *wind_speed*≤ 1.1, *min_temp*> -12.5, *WDI*≤ 1.19134, *pressure*≤1018.5
*tr*_M3	1	2.15695 (142.37)	10	*min_temp*≤ 7.5, *wind_speed*≤ 2.45, *wind_speed*≤ 1.1, *min_temp*≤-5.5, *cloud_cover*≤ 0.29
*tr*_M3	4	2.1315 (118.88)	10	*min_temp*≤ 7.5, *wind_speed*≤ 2.45, *wind_speed*≤ 1.1, min_temp >-5.5, humidity ≤ 0.725, cloud_cover > 0.19
***tr*_M4**	**1**	2.16811 (**154.73**)	**5**	***min_temp*≤ 7.5, *wind_speed*≤ 2.45, *wind_speed*≤ 1.1, *min_temp*≤ -12.5**
*tr*_M4	11	2.1558 (141.17)	6	*min_temp*≤ 7.5, *wind_speed*≤ 2.45, *wind_speed*≤ 1.1, *min_temp*>-12.5, *cloud_cover*≤ 0.685, *WDI*> 1.19134, *humidity*≤ 0.715, *year_month*> 0.92

*) Retransformed means from the models of *tr_PM*10 are given in brackets.

The plot of predicted values of PM10 by the model tr_M4 is given in Figure 4. Comparison with actual measured concentrations of PM10 in the upper section of Figure 2 shows very good agreement.

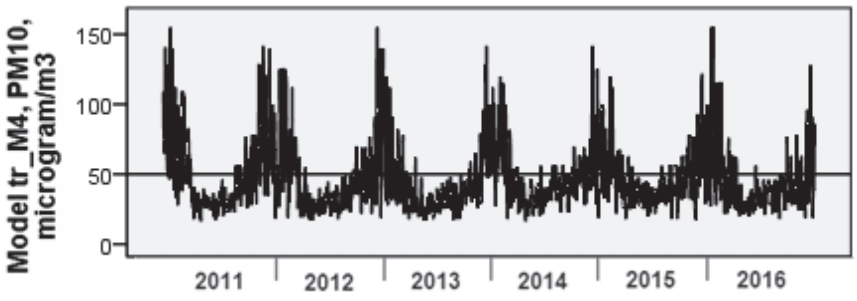

FIGURE 4. Sequence plot of predicted values of PM10, obtained by the CART model *tr*_M4

APPLICATION OF THE MODELS FOR SHORT-TERM FORECASTS

Using the models with the data until 31 December 2016 the PM10 concentrations for 2 days (1 and January 2017) were forecasted. In Figure 5 illustrates the predicted values obtained using the four CART models over the last 5 days - from December 27, 2016 to December 31, 2016 (to the left of the vertical line), and the forecasts for the next two days – 1 and 2 January 2017 (to the right of the vertical line). The comparison shows that the *tr*_M3 and *tr*_M4 models are much more accurately fitted to the measured daily average values of *PM*10. The horizontal lines represent the permissible daily upper limit of 50 micrograms per cubic meter.

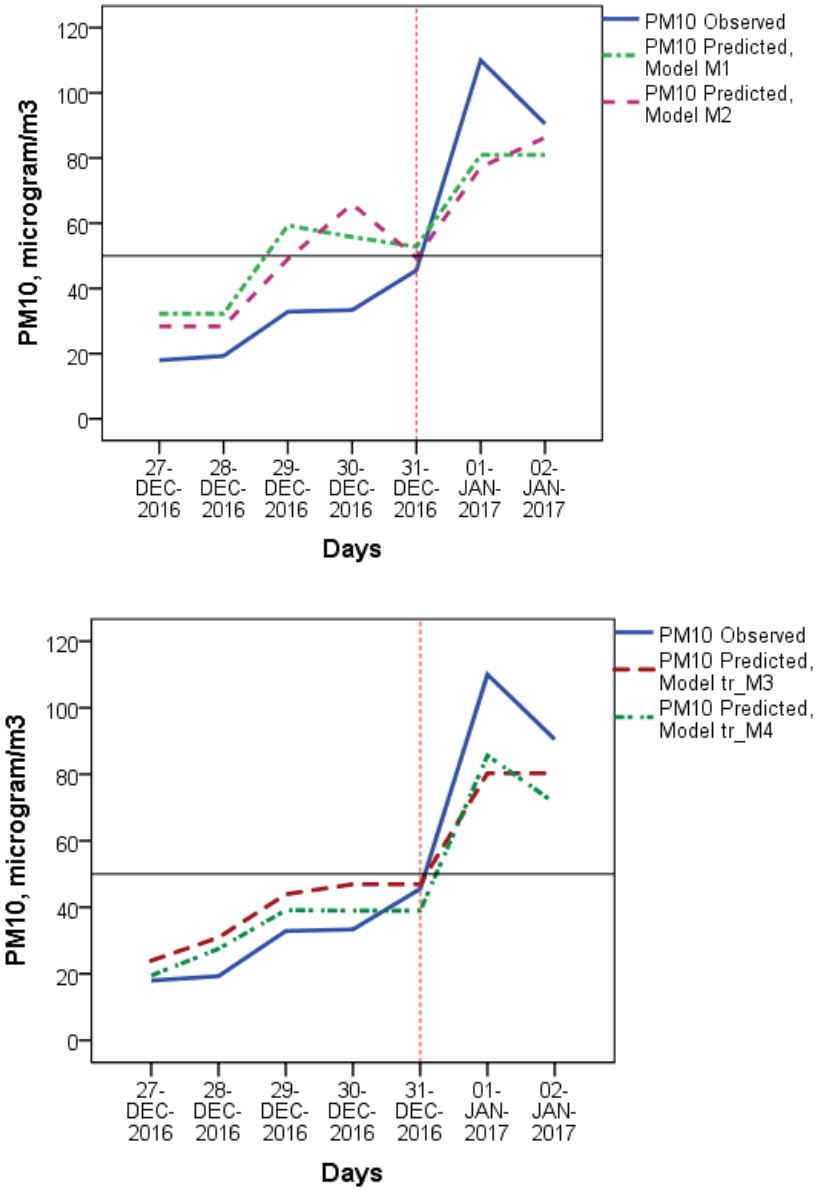

FIGURE 5. Comparison of measured and predicted *PM*10 concentrations

CONCLUSION

The data from measurements of average daily PM10 concentration were examined, which is a problematic air pollutant in many Bulgarian cities. As an example, the data for the city of Pleven, located in the Danube plain, north Bulgaria, are processed and analyzed. The investigated data cover a period of 6 years, between 2011 and 2016. It has been found that over this period regulatory air pollution health limits exceeded constantly. Modeling is performed using a powerful and flexible data mining statistical technique – classification and regression trees (CART). Four optimal CART models for modeling PM10 levels depending on 8 meteorological variables and one time variables are obtained, analyzed, and compared. Two of the models are built for the transformation of variable *tr_PM*10, which has almost normal distribution. Very good fit up to 78% is achieved compared with the experiment. An application of the models for forecasting pollution for the next 2 days is illustrated using the measurements used. The influence of individual variables on the models has been determined. The proposed approach shows significant advantages for the investigation of relationships between time series, and particularly the adverse impact on public health of air pollution with the goal of forecasting, prevention, and control of air quality in urban areas.

The conducted analysis is an alternative to the official reports by the regional environmental and water inspectorate – Pleven and an independent investigation of the measured concentrations of PM10.

ACKNOWLEDGEMENTS

This work was supported by Paisii Hilendarski University of Plovdiv NPD grant MU17-FMI-003.

REFERENCES

1. P. T. Nastos, A. G. Paliatsos, M. B. Anthracopoulos, E. S. Roma, and K. N. Priftis (2010) Outdoor particulate matter and childhood asthma admissions in Athens, Greece: a time-series study, *Environmental Health* **9**(1), Art. Number 45, doi: 10.1186/1476-069X-9-45.
2. K. Gass, M. Klein, H. H. Chang, W. D. Flanders, and M. J. Strickland (2014) Classification and regression trees for epidemiologic research: an air pollution example, *Environmental Health* **13**, Art. Number 17.
3. *Air* quality in Europe - 2014 report, European Environment Agency Publications, 19 Nov 2014, Accessed 10 June 2017, Available at http://www.eea.europa.eu/publications/air-quality-in-europe-2014/at_download/file.
4. Executive Environment Agency (ExEA), Bulgaria, http://eea.government.bg/en, Accessed 10 June 2017.
5. L. Jian, Y. Zhao, Y. P. Zhu, M. B. Zhang, and D. Bertolatti (2012) An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, *China, Sci. Total Environ.* **426**, 336–345, doi: 10.1016/j.scitotenv.2012.03.025.
6. P.W.G. Liu (2009) Simulation of the daily average PM10 concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis, *Atmos. Environ.* **43**, 2104-2113, doi: 10.1016/j.atmosenv.2009.01.055.
7. M.L. Martın, I. J. Turias, F. J. Gonzalez, P. L. Galindo, F. J. Trujillo, C. G. Puntonet, and J. M. Gorriz (2008) Prediction of CO maximum ground level concentrations in the Bay of Algeciras, Spain using artificial neural networks, *Chemosphere* **70**(7), 1190–1195.
8. A. Prakash, U. Kumar, K. Kumar, and V. Jain (2011) A wavelet-based neural network model to predict ambient air pollutants' concentration, *Environ. Model. Assess.* **16**(5), 503-517.
9. A.V. Ivanov and S.G. Gocheva-Ilieva. "Short-time particulate matter PM10 forecasts using predictive modeling techniques." in *AMiTaNS'13*, AIP Conference Proceedings 1561, edited by M.D. Todorov (American Institute of Physics, Melville, NY, 2013), pp. 209-218, doi:10.1063/1.4827230.
10. T. Slini, A. Kaprara, K. Karatzas, and N. Moussiopoulos (2006) PM10 forecasting for Thessaloniki, Greece, *Environ. Modell. Softw.* **21**(4), 559-565, doi: 10.1016/j.envsoft.2004.06.011.
11. J. Neto, F. Ferreira, P. M. Torres, and F. Boavida (2009) Lisbon air quality forecast using statistical methods, *Int. J. Environment and Pollution* **39**(3/4), 333–339.
12. Salford Systems Data Mining and Predictive Analytics Software Modeler, SPM Version 8.0 (Salford Systems, San Diego, 2016).
13. SPSS IBM Statistics, http://www-01.ibm.com/software/analytics/spss/, Accessed 10 June 2017.

14. 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, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:152:0001:0044:EN:PDF, Accessed 10 June 2017.
15. Air Quality Standards. European Commission. Environment. 2015, http://ec.europa.eu/environment/air/quality/standards.htm, Accessed 10 June 2017.
16. Regional Inspectorate of Environment and Water - Pleven, Reports on the state of the environment, 2011-2016, http://riew-pleven.eu/doc/docladiOS/,Accessed 10 June 2017. [in Bulgarian]
17. I. K. Yeo and R. A. Johnson (2000) A new family of power transformations to improve normality or symmetry, *Biometrika* **87**(4), 954–959, doi:10.1093/biomet/87.4.954.
18. L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees* (Wadsworth Intern, Belmont, 1984).
19. R. Nisbet, J. Elder, and G. Miner, *Handbook of Statistical Analysis and Data Mining Applications* (Elsevier Academic Press, Burlington, 2009).
20. A. J. Izenman, *Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning*(Springer, New York, 2008).
21. D. Steinberg and P. Colla, *CART: Tree-Structured Non-Parametric Data Analysis* (Salford Systems, San Diego, 1995).
22. CART® Classification and Regression Trees, Salford Systems, San Diego, 2012, http://www.salford-systems.com/en/products/cart, Accessed 10 June 2017.
23. MARS 3.0, Technical guide, Salford Systems, San Diego, 2011.