



Modeling of PM₁₀ Air Pollution in Urban Environment Using MARS

Snezhana G. Gocheva-Ilieva^(✉), Atanas V. Ivanov, Desislava S. Voynikova,
and Maya P. Stoimenova

University of Plovdiv Paisii Hilendarski, 24 Tsar Asen Street, 4000 Plovdiv, Bulgaria
{snow,aivanov,desi_voynikova,mstoimenova}@uni-plovdiv.bg

Abstract. In the modern world, attention is increasingly drawn to the pressing problem of atmospheric air pollution, which is a serious threat to human health. Worldwide, China, India, Indonesia and some of the countries in Europe, including Bulgaria, are the most polluted countries. To help solve these issues, a very large number of scientific studies have been devoted, including the study, analysis and forecasting of atmospheric air pollution with particulate matter PM₁₀. In this study the PM₁₀ concentrations in the town of Smolyan, Bulgaria are examined and mathematical models with high performance for prediction and forecasting depending on weather conditions are developed. For this purpose, the powerful method of multivariate adaptive regression splines (MARS) is implemented. The examined data cover a period of 9 years - from 2010 to 2018, on a daily basis. As independent variables, 7 meteorological factors are used - minimum and maximum daily temperatures, wind speed and direction, atmospheric pressure, etc. Additional predictors also used are lagged PM₁₀ and meteorological variables with a delay of 1 day. Three time variables are included to account for time. Multiple models are created with interactions between predictors up to the 4th order. The obtained best MARS models fit to over 80% of measured data. The models are used to forecast PM₁₀ concentrations for 7 days ahead of time. This approach could be applied for real predictions and development of computer and mobile applications.

Keywords: Air pollution · Particulate matter PM₁₀ · Multivariate adaptive regression splines (MARS) · Forecasting

1 Introduction

Worldwide, attention is increasingly drawn to the quality of atmospheric air and the negative impact of air pollutants on the human health. In the field of ecology, there are numerous scientific studies, which indicate the worrying statistics related to bad air quality and the growing number of people who suffer

This work has been accomplished with the financial support of the MES by the Grant No. D01-221/03.12.2018 for NCDSC, part of the Bulgarian National Roadmap on RIs.

© Springer Nature Switzerland AG 2020

I. Lirkov and S. Margenov (Eds.): LSSC 2019, LNCS 11958, pp. 237–244, 2020.

https://doi.org/10.1007/978-3-030-41032-2_27

from disease and die as the result of air pollution [1,2]. Recent data show that Bulgaria ranks among the European countries with the most polluted air and the highest levels of particulate matter $10\ \mu\text{m}$ in diameter or smaller (PM10) [3,4]. The main reasons, which facilitate the generation of PM10 are the emissions of various fuel-based commercial and domestic processes, busy traffic and other, predominantly during the winter periods. In Bulgaria, 36 automated measuring stations constantly monitor the levels of 12 main pollutants, including PM10, and are controlled by the national Executive Environment Agency (EExA) [5].

In the literature, various methods for investigation and forecasting of air pollutant concentrations are used in order to ensure air quality and to prevent pollution. Multivariate linear and non-linear regression, principal component analysis, factor analysis, cluster analysis, and other have a wide range of applications [6,7]. In a large number of publications, stochastic methods for modeling time series of air pollutants are applied (see [8,9] and the literature cited therein). The most frequently used methods of this type are ARIMA, transfer function methods and others, based on the methodology of Box-Jenkins [10].

In recent years, alongside these methods, the scientific community is increasingly more interested in new computer-oriented machine-learning data-driven methods. These methods extract direct dependencies and generate mathematical models from the data. The most widely-used methods are: Artificial neural networks (ANN), Support vector machine (SVM) regression, Random Forest (RF), Multivariate adaptive regression splines (MARS), etc. In [11] forecasts are made for the concentrations of SO₂ by applying three different methods - least square SVR, MARS, and M5 Model Tree. It was found that the MARS method is the second best for predicting SO₂ concentrations. Authors of [12] present predicted results for PM10 concentrations for the city of Sarajevo, obtained using the ANN method and input data for meteorological variables (wind speed, humidity, temperature and pressure) over the period from 2010 to 2013. Predictive RF models are built and analyzed in [13] for data from 31 air quality monitoring sites in Switzerland using surface meteorological, synoptic scale, boundary layer height, and time variables to explain daily PM10 concentrations. A recent summary paper [14] presents the current state of air pollution forecasting methods and published results.

This study applies the non-parametric data-driven regression technique MARS to analyze and model the concentrations of PM10 in the town of Smolyan, Bulgaria. The objective of the developed empirical case study is to construct the most suitable models, which describe the measured data with the highest degree of predictive power and to forecast the future pollution for up to 7 days ahead.

Statistical procedures are carried out using the software package Salford Predictive Modeler (SPM) [15] and IBM SPSS [16].

2 Description of the Study Area and Data

The town of Smolyan is located in South Bulgaria and is the administrative center of Smolyan municipality and Smolyan province. The town has a population

of 28,000 people and is the smallest main provincial town in Bulgaria. The climate of Smolyan is transitional-continental with predominant Aegean influence. The terrain is mountainous with an abundance of coniferous forests. The average altitude is around 1000 m. Winters are mild and snowy with a stable snow-cover between 80 and 120 days, and summers are cool and humid. The average temperature in January is between -1 and -5 °C, and in July—between 13 and 20°. The town of Smolyan was chosen for this study as representative of mountain settlements, regional centers in Bulgaria, with systemic pollution of PM10. Similar examples are the cities of Blagoevgrad, Kardzhali and others.

The investigation is conducted using average daily measurements of the pollutant PM10 ($\mu\text{g}/\text{m}^3$) in the town of Smolyan over a period of 9 years from 1 January 2010 to 30 November 2018 or the total number of observations is $N = 3254$. The missing data for the observed time interval of the PM10 variable are 167, which is about 5%. In the modeling procedures, the missing data are replaced using linear interpolation. The analysis also includes 13 initial predictors: CLOUD - cloud cover, HUMID - relative air humidity, MAXT - maximum daily temperature, MINT - minimum daily temperature, PRESS - atmospheric pressure, WINDSP - winds speed, PRECIP - precipitation, as well as the lagged variables: PM10_1, MINT_1, and others with the values from the previous day. In order to maintain correspondence with the weather, the models also use 3 time variables: YEAR_MONTH, MONTH, and MONTH_DAY.

Table 1 shows descriptive statistics of the initial data for the town of Smolyan. The table indicates that the average value of the pollutant PM10 is $44.14 \mu\text{g}/\text{m}^3$, which is close to the maximum permissible average daily threshold of $50 \mu\text{g}/\text{m}^3$ as per European and national standards [17, 18]. The maximum recorded value is $260.5 \mu\text{g}/\text{m}^3$, which exceeds the threshold by a factor of five. The initial analysis clearly shows that the PM10 air pollutant is problematic for the town.

Figure 1 shows the sequence plot of the time series for the PM10 pollutant over the considered 9-year period. The horizontal line in the plot represents the regulatory requirements of $50 \mu\text{g}/\text{m}^3$. It clearly outlines multiple instances where this threshold value is exceeded. The exceedances for the pollutant are observed

Table 1. Descriptive statistics of the initial data.

Variable	Mean	Median	Std. dev.	Skewness	Kurtosis	Minimum	Maximum
PM10, $\mu\text{g}/\text{m}^3$	44.14	33.48	29.59	1.858	4.334	1.64	260.5
MINT, °C	8.78	9.00	9.745	-0.260	-0.592	-23	29
MAXT, °C	14.19	15.00	8.984	-0.263	-0.784	-15	33
PRESS, mbar	1016	1016	6.864	0.211	0.315	990	1039
HUMID, %	0.767	0.77	0.131	-0.260	-0.694	0.28	1
CLOUD, %	0.345	0.27	0.270	0.727	-0.478	0	1
PRECIP, cm	4.781	0.80	9.00	3.874	24.833	0	126.1
WINDSP, m/s	6.808	6.00	3.179	1.459	3.639	1	28

Note. Std. Err. of Skewness for all variables is 0.044 and Std. Err. of Kurtosis is 0.088.

mainly during the cold months of the year. This is explained by the fact that households use mainly solid fuel and coal for heating.

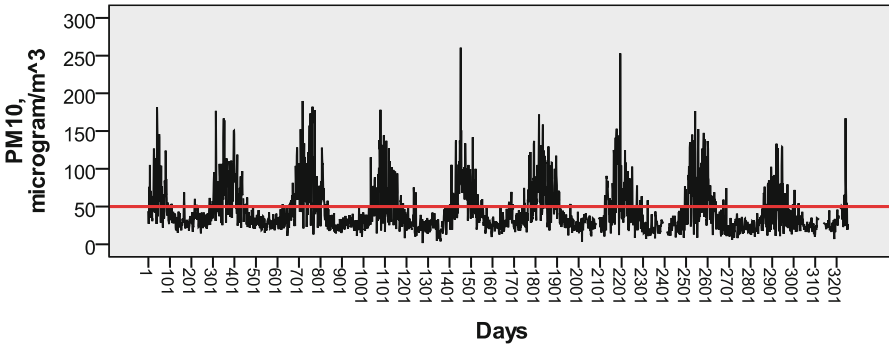


Fig. 1. Sequence plot of the measured daily concentrations of PM10 for the town of Smolyan.

3 MARS Method

The MARS method was developed by the American statistician and physicist Friedman in 1990–1991 [19]. The advantage of the MARS method lies in its ability to process both small- and large-scale datasets. If the dependent variable (here PM10) is $y = y(X)$ and the predictors $X = (X_1, X_2, \dots, X_p)$ are vectors with dimension N , the MARS model $\hat{y} = \hat{y}_{[M]}$ has the following form:

$$\hat{y}_{[M]} = b_0 + \sum_{j=1}^M b_j BF_j(X), \tag{1}$$

where $b_0, b_j, j = 1, 2, \dots, M$ are the coefficients in the model, $BF_j(X)$ are its basis functions (BFs), M is their number, p is the number of predictors in the model. The form of one-dimensional BFs is one of the follows

$$BF_j(X) = \max_{X_k}(0, X_k - c_{k,j}), \quad BF_j(X) = \max_{X_k}(c_{k,j} - X_k, 0), \tag{2}$$

where $c_{k,j} \in X_k$ are nodes selected and fixed by the MARS algorithm, X_k is a variable (predictor), $k = 1, 2, \dots, p$. For the non-linear interactions BFs are built as products of other BFs.

When constructing the model, some control parameters need to be set by the researcher. The maximum number of basis functions (L_0) and the maximum r of multipliers (i.e. degree of interactions) in BFs are set, where $L_0 \geq 3p$. The algorithm consists of two steps. During the first step, initial value for b_0 is taken (for example $b_0 = \min_i y_i$) and consequently the model is complemented by

BFs of type (2). For the current model with L number of BFs, the selection of variables and nodes is determined by the condition for minimizing the sum

$$SSL_{[L]} = \sum_{i=1}^N [y_i - \hat{y}_{[L],i}(X)]^2. \quad (3)$$

The second step of the method includes an assessment of the accuracy of the constructed MARS model and reduction of the number of BFs using the general cross-validation (GCV) criterion, which takes into account not only the residuals error but also the overall error of the model [19]. BFs which do not contribute significantly to the accuracy of the model are removed. The form of the criterion is:

$$GCV_{[m]} = \frac{\sum_{i=1}^N [y_i - \hat{y}_{[m],i}(X)]^2}{N(1 - C(n)/N)^2}, \quad m = 1, 2, \dots, L_1, \quad L_1 \leq L_0. \quad (4)$$

Here $C(m) = m + \delta(m - 1)/2 \in [2, 3]$. The second step of the algorithm prevents the overfitting of the model. The best MARS models were selected with the highest coefficient of determination R^2 , the highest GCV R^2 , obtained after step 2, and the minimum values of the root mean square error (RMSE), given by the expressions

$$R^2 = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (5)$$

where y_i are the observed values of the dependent variable $y = PM10$, \bar{y} is its mean, \hat{y}_i are the predicted values by the MARS model.

4 Results and Discussion

4.1 Construction of MARS Models

An important part in constructing MARS models is the choice of control parameters. With a relatively large volume N of the sample, a maximum number of BFs of 100, 200 and 300, and a maximum degree of interactions between the predictors $r = 2, 3$ and 4 were selected. Furthermore, a more detailed analysis of the PM10 observations showed only two very high measured values - $260.5 \mu\text{g}/\text{m}^3$ and $253 \mu\text{g}/\text{m}^3$, respectively. In order not to unnecessarily complicate the models, these two outliers were replaced by linear interpolation.

The conducted model procedures showed that the best predictive results could be obtained with the use of all 7 meteorological variables from Table 1 and their lagged variables and lagged variable PM10_1, taking into account the PM10 pollution from the previous day. As predictors, the following time variables were also used: MONTH, YEAR_MONTH and MONTH_DAY respectively, given by the expressions:

$$YEAR_MONTH = YEAR + (MONTH - 0.5)/12, \quad (6)$$

$$MONTH_DAY = MONTH + (DAY - 0.5)/31 \tag{7}$$

Interactions between these time variables was ruled out. In order to assess the future forecasting model capabilities, the data for the last 7 days of the sample (24–30 November 2018) were not included in the modeling process and were used as holdout data. The total maximum number of predictors is 18. Table 2 shows the main statistics of selected best MARS models.

Table 2. Summary statistics of the selected best MARS models of PM10.

MARS model ($L_0 r$)	Number of BFs	$R^2, \%$	$GCV R^2$	RMSE
M1 (200; 2)	112	81.34	78.21	12.460
M2 (200; 3)	138	84.23	80.00	11.497
M3 (200; 4)	152	86.66	82.86	10.699
M4 (300; 2)	161	83.64	78.13	11.711
M5 (300; 3)	179	86.32	81.35	10.707

The obtained best MARS models describe the observed PM10 data with $R^2 = 86\%$ and $GCV R^2 = 83\%$. Models M3 (200; 4) and M5 (300; 3) have the highest performance, followed by M2 (200; 3). Since their respective RMSEs are not significantly different, we choose models M2 and M5 as the simplest models with interaction between predictors up to 3th degree. Figure 2 shows the correspondence between M5 and PM10 with a 5% confidence interval.

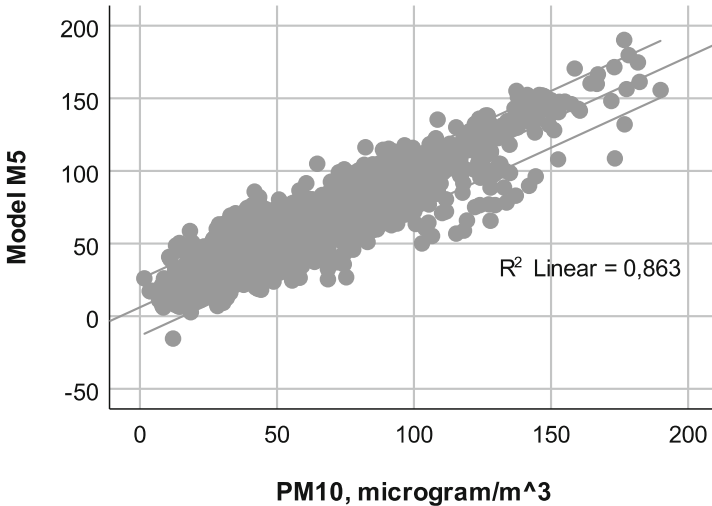


Fig. 2. Comparison of the predicted PM10 values using model M5 and the measured PM10 values.

4.2 Application of Models for Forecasting Future Pollution

The MARS method has the advantage over many other methods that it generates flexible models which allow to obtain very easy the predictions and forecasts through formulas (1)–(2) at given predictor values. Figure 3 illustrates the calculated estimates of M2 and M5 models compared to PM10 data for 7 days ahead. Very good correspondence is observed.

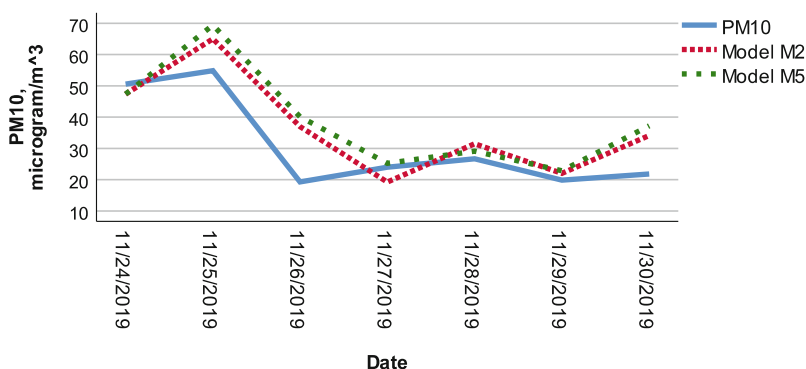


Fig. 3. Forecasting results obtained using MARS models M2 and M5 against the measured PM10 values for 7 days ahead.

5 Conclusion

With the help of the powerful data-driven method MARS statistical models for analyzing and forecasting of PM10 atmospheric air pollution data, depending on changes in weather conditions are built. Models with very good statistical characteristics were obtained. The agreement of models with measured data and their high performance for predicting future pollution concentrations has been demonstrated.

The performed analyzes and obtained results suggest that MARS can successfully be used to model other air pollutants as well as similar time series of ecology data.

References

1. Maji, J.K., Dikshit, A.K., Deshpande, A.: Disability - adjusted life years and economic cost assessment of the health effects related to PM2.5 and PM10 pollution in Mumbai and Delhi, in India from 1991 to 2015. *Environ. Sci. Poll. Res.* **24**(5), 4709–4730 (2017)

2. Wang, C., Zhou, X., Chen, R., Duan, X., Kuang, X., Kan, H.: Estimation of the effects of ambient air pollution on life expectancy of urban residents in China. *Atmos. Environ.* **80**, 347–351 (2013)
3. Piepoli, M., et al.: 2016 European Guidelines on cardiovascular disease prevention in clinical practice: the Sixth Joint Task Force of the European Society of Cardiology and Societies on cardiovascular disease prevention in clinical practice. *Eur. Heart J.* **37**(29), 2315–2381 (2016)
4. European Environment Agency, Air quality in Europe - 2018 report. <https://www.eea.europa.eu/publications/air-quality-in-europe-2018>. Accessed 22 Feb 2019
5. Executive Environment Agency. <http://eea.government.bg/en/>. Accessed 22 Feb 2019
6. Stadlober, E., Hubnerova, Z., Michalek, J., Kolar, M.: Forecasting of daily PM10 concentrations in Brno and Graz by different regression approaches. *Austrian J. Stat.* **41**(4), 287–310 (2012)
7. Ng, K.Y., Awang, N.: Multiple linear regression and regression with time series error models in forecasting PM10 concentrations in Peninsular Malaysia. *Environ. Monit. Assess.* **190**, 63 (2018)
8. Zheleva, I., Veleva, E., Filipova, M.: Analysis and modeling of daily air pollutants in the city of Ruse, Bulgaria. In: Todorov, M. (ed.) AIP Conference Proceedings, vol. 1895, no. 1, p. 030007. American Institute of Physics, Melville (2017)
9. Gocheva-Ilieva, S.G., Ivanov, A.V., Voynikova, D.S., Boyadzhiev, D.T.: Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach. *Stoch. Environ. Res. Risk Assess.* **28**(4), 1045–1060 (2014)
10. Box, G.E.P., Jenkins, G.M., Reinsel, G.S.: *Time Series Analysis, Forecasting and Control*, 3rd edn. Prentice-Hall Inc., New Jersey (1994)
11. Kisi, O., Parmar, K.S., Soni, K., Demir, V.: Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, and M5 model tree models. *Air Qual. Atmos. Health* **10**(7), 873–883 (2017)
12. Dedovic, M.M., Avdakovic, S., Turkovic, I., Dautbasic, N., Konjic, T.: Forecasting PM10 concentrations using neural networks and system for improving air quality. In: *Proceedings of the 11th International Symposium Telecommunications (BIH-TEL)*, Article no. 7775721. IEEE (2016)
13. Grange, S.K., Carslaw, D.C., Lewis, A.C., Boleti, E., Hueglin, C.: Random forest meteorological normalisation models for Swiss PM10 trend analysis. *Atmos. Chem. Phys.* **18**(9), 6223–6239 (2018)
14. Bai, L., Wang, J., Ma, X., Lu, H.: Air pollution forecasts: an overview. *Int. J. Environ. Res. Public Health* **15**(4), 780 (2018)
15. Salford Predictive Modeler 8. <https://www.salford-systems.com>. Accessed 22 Feb 2019
16. SPSS IBM 25. <https://www.ibm.com/products/>. Accessed 22 Feb 2019
17. 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)
18. European Commission, Environment, Air, Air Quality Standards (2018). <http://ec.europa.eu/environment/air/quality/standards.htm>. Accessed 22 Feb 2019
19. Friedman, J.H.: Multivariate adaptive regression splines (with discussion). *Ann. Stat.* **19**(1), 1–141 (1991)