Neural Computing and Applications (2019) 31:9023–9039
https://doi.org/10.1007/s00521-019-04432-1

ORIGINAL ARTICLE

Regression trees modeling of time series for air pollution analysis and forecasting

Snezhana Georgieva Gocheva-Ilieva[1] · Desislava Stoyanova Voynikova[1] · Maya Plamenova Stoimenova[1] · Atanas Valev Ivanov[1] · Iliycho Petkov Iliev[2]

Received: 13 October 2017 / Accepted: 7 August 2019 / Published online: 17 August 2019

Abstract
Solving the problems related to air pollution is crucial for human health and the ecosystems in many urban areas throughout the world. The accumulation of large arrays of data with measurements of various air pollutants makes it possible to analyze these in order to predict and control pollution. This study presents a common approach for building quality nonlinear models of environmental time series by using the powerful data mining technique of classification and regression trees (CART). Predictors for modeling are time series with meteorological, atmospheric or other data, date-time variables and lagged variables of the dependent variable and predictors, involved as groups. The proposed approach is tested in empirical studies of the daily average concentrations of atmospheric PM10 (particulate matter 10 μm in diameter) in the cities of Ruse and Pernik, Bulgaria. A 1-day-ahead forecasts are obtained. All models are cross-validated against overfitting. The best models are selected using goodness-of-fit measures, such as root-mean-square error and coefficient of determination. Relative importance of the predictors and predictor groups is obtained and interpreted. The CART models are compared with the corresponding models built by using ARIMA transfer function methodology, and the superiority of CART over ARIMA is demonstrated. The practical applicability of the models is assessed using 2×2 contingency tables. The results show that CART models fit well the data and correctly predict about 90% of measured values of PM10 with respect to the average daily European threshold value of 50 $\mu g/m^3$.

Keywords Air pollution modeling · Time series · Classification and regression trees (CART) · Pollution forecast

Mathematics Subject Classification 62M10 · 62M20 · 62P12

1 Introduction

The processes of atmospheric air pollution in urban areas are complex and depend on a large number of factors, such as meteorological and atmospheric influence, chemical reactions, industrial, transport or household activities, etc. When the permissible levels are exceeded, this type of pollution is extremely harmful for the health of the population and causes environmental damage. Measuring, studying, preventing and controlling these require continuous effort in many countries around the world. In the European Union, air quality, the permissible threshold for concentrations of the main pollutants, the measures and standards are regulated by official documents [1, 2]. The systematic measurement of the main air components, including carbon oxides, nitrogen oxides, ozone, particulate matter, etc., leads to an accumulation of large data arrays in the form of time series. This allows for air quality modeling and forecasting in order to derive significant information, define classifications and reveal dependencies. Building the models allows for the investigation of pollution processes over time or in space, the degree of influence of various factors or groups of factors (meteorological, environmental, atmospheric, etc.) on these processes, as well as to prepare forecasts for expected pollution levels, usable for prevention and control. It should be noted that

✉ Snezhana Georgieva Gocheva-Ilieva
snegocheva@gmail.com; snow@uni-plovdiv.bg

[1] Department of Applied Mathematics and Modeling, Faculty of Mathematics and Informatics, University of Plovdiv Paisii Hilendarski, 24 Tzar Asen St, 4000 Plovdiv, Bulgaria

[2] Department of Physics, Technical University – Sofia, Branch Plovdiv, 25 Tzanko Djusstabanov St, 4000 Plovdiv, Bulgaria

this type of data could be characterized by nonlinear behavior and different type of dependencies, including autoregressive and local in time and space, which make the modeling process more difficult and require the selection of suitable method.

Among the methods for stochastic time series modeling in this field, the Box–Jenkins ARIMA method [3] and its various versions as univariate, vector or transfer function are widely popular. ARIMA models allow for a good description of linear relationships within the time series, which in many cases provides satisfactory results. ARIMA modeling and principal component analysis are applied in [4] for simulation and forecasting of the daily average PM10 concentrations. In [5], ARIMA models are developed and analyzed for six air pollutants in the town of Ploiesti, Romania. The obtained results demonstrate unsatisfactory performance for some of the models, depending on the data behavior. Univariate ARIMA models are constructed and applied for short-term forecast of PM10 air pollution in [6, 7]. Often, the performance of ARIMA is combined and compared against other methods. For example, a hybrid ARIMA–ANN is developed in [8], where the linear part of the time series is modeled using ARIMA, and the nonlinear one using ANN. A combination of seasonal ARIMA and support vector machine (SVM) approach is employed for PM10 and ozone modeling in [9]. Another recent example is the study [10], where the authors use and compare the capabilities of 4 methods—univariate and vector ARIMA, multilayer perceptron (MLP) and SVM with regression for forecasting monthly averaged PM10 concentration in Oviedo, Spain, for a period of 7 years. It is found that SVM model overperforms other models in forecasting 1 month ahead and also for the following 7 months. In [11], wavelet analysis and wavelet ARIMA models are presented.

Although ARIMA is a flexible method for modeling both the deterministic part of the time series (linear autoregression processes) and the stochastic part (moving average processes), as well as trends, it lacks the ability to grasp well the nonlinear and local dependencies. For this reason, along with the standard methods of multidimensional statistical analysis, powerful and adaptive predictive methods and techniques for the study of time series in ecology based on artificial intelligence and data mining are adapted. Of these methods applied to time series, the most actively used are neural networks (NN), MLP, wavelet analysis, SVM, CART, MARS, fuzzy functions and more, as well as various combinations and comparisons of them and with other methods.

Frequently, air pollution is modeled by applying neural networks (NN), MLP, fuzzy logic, SVR. In [12], daily averaged PM10 and PM2.5 concentrations are forecasted using a recursive artificial NN, a feed-forward NN and a multiple linear regression with meteorological variables. Authors of [13] applied principal component analysis (PCA), ANN and *k*-means clustering technique in order to forecast PM10 and PM2.5 concentrations for the next day. A hybrid computational intelligence system based on ensembles learning classifiers as feed-forward NN, fuzzy logic and random forest is presented in [14] for modeling CO, NO, NO_2, SO_2 and ozone O_3 pollution levels. Efficacy of MLP, SVR, wavelet analysis and other methods on artificial intelligence is used in recent papers [9–11, 15, 16].

Actively used for air pollutant modeling and forecasting are the nonparametric data-driven data mining methods: CART, multivariate adaptive regression splines (MARS) and others. Decision trees were employed since the mid-1960s (see for instance Morgan and Sonquist [17]) and later developed further by Bremein et al. [18] as CART in 1984. The tree-based CART models are applied in [19] to analyze and predict the maximum daily concentration of a ground ozone (O_3) in three cities, Canada. The paper [20] investigates the concentration of PM10 in Thessaloniki, Greece, over a period of 7 years in relation to meteorological and other series using and comparing multivariate linear regression, PCA, NN and CART. The authors of [21] apply and compare four machine learning methods for predicting PM10 exceedances over the European upper daily limit of 50 $\mu g/m^3$ in Helsinki: logistic regression, decision trees, MARS and NN. However, in the same study, probably due to the small sample sizes and the specifics of predictors, CART showed lower performance compared to other methods. Recent applications of CART and boosted regression trees for studying air pollutants depending on meteorological conditions, road traffic, etc., have been presented in [22, 23]. Another application is [24], where these are built and used for short-term predictive exploratory CART models for PM10 in the town of Pleven, Bulgaria, by using nonlagged meteorological variables and date-time series as predictors.

One of the first studies on time series with the powerful technique MARS is [25] where the method and simulations are presented to demonstrate its abilities for capturing of nonlinear local dependences in data. Based on modern optimization techniques, allowing to improve model accuracy and performance, new versions of classical MARS were developed, namely conic MARS (CMARS), robust conic MARS (RCMARS), robust MARS (RMARS), conic generalized partial linear modeling (CGPLM with (C)MARS), robust CGPLM (RCGPLM), etc. [26–29]. Powerful applications of these methods for time series, including studies in environmental sciences, are reported in [30–34]. In [30], high-quality RCMARS models for precipitation data are presented and are compared with corresponding MARS and CMARS models. Authors of [31] applied CMARS method for solving inverse problems

regarding atmosphere reflection values. Recent papers using and comparing MARS and CART are [35, 36].

A detailed analysis of the results in the field of air pollution modeling and the most commonly used methods can be found in the recent review papers [37, 38], with [37] posing an emphasis on PM10.

This paper explores the capabilities of the powerful data mining CART technique for statistical modeling of datasets from measurements of air pollutants and similar type of environmental data. The goal is to present a common approach for building quality nonlinear models in order to describe and analyze relevant time series and to extract essential information from them, which helps to study the causes of pollution, forecasting and control of air quality in urban areas. In particular, the goal is to demonstrate the capabilities of CART for modeling similar levels of the air pollutant PM10 by classifying these in subsets with respect to combinations of meteorological and other conditions, taking into account autoregression dependencies and date-time variables. The predictors will be involved in the modeling procedure by groups. An important task is to establish the relative importance of predictor groups in the models and conduct the interpretation of the model results according to the actual state of air pollution. This is essential for the quality of forecasts for future pollution. In this study, the goal is to forecast PM10 concentrations for a 1 day ahead based on average daily observations. A specific objective is to compare the quality of the obtained models for forecasting PM10 1 day ahead in three cases: (i) by considering known and/or forecasted values of the meteorological variables at time $t + 1$ for predicting PM10 at time $t + 1$; (ii) by considering the meteorological variables at time t (lagged variables) for predicting PM10 at time $t + 1$; (iii) by considering simultaneously meteorological variables both at time t and $t + 1$ for predicting PM10 at time $t + 1$. Special attention will be given to compare the performance of the CART models with the results from other methods widely used in environmental statistics. In our case, we chose the ARIMA transfer function methodology to demonstrate capabilities of CART.

The paper is organized as follows: (1) problem setup and description of the proposed CART approach; (2) modeling stages; (3) empirical results from applying CART on data for particulate matter PM10 pollution in two Bulgarian cities—Ruse and Pernik; (4) analysis of accuracy and adequacy of models and their quality for forecasting; (5) assessment of the practical application of the models in accordance with the European PM10 standards; (6) comparison of the main results with ARIMA models for the same time series; (7) discussion and conclusion.

Computer simulations are performed using Salford Predictive Modeler, SPSS software and Wolfram Mathematica [39–41].

2 CART methodology for time series modeling

2.1 Problem formulation

It is assumed that the measurements data for the concentrations of a given air pollutant Y are known as a time series as well as the respective time series $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ for meteorological, atmospheric, etc., variables, which influence it. We will consider that $\mathbf{X}$ also includes a group of date-time variables. Let the time increment values be denoted by $t = 1, 2, 3, \ldots, n$, where n is the total number of observations in the series, Y_t and $\mathbf{X}_t = (X_{1,t}, X_{2,t}, \ldots, X_{m,t})$ are values of the series at the moment t.

We set out to build a model which finds an explicit form of the dependence of Y on the predictors $\mathbf{X}$, making it possible to predict the value of the pollutant. In the general case, the series of the considered type may not be stationary, i.e., their probability distribution may depend on time, and the mean value and variance may not be constant. Considering such specifics, their modeling is often not trivial.

We will consider the following general kind of dependence

$$Y_t = f\Big(Y_{t-1}, \ldots, Y_{t-p'}, X_{1,t}, \ldots, X_{m,t}, X_{1,t-1}, \ldots, X_{1,t-q_1'} \ldots, X_{m,t-1}, \ldots, X_{m,t-q_m'}\Big) + \varepsilon_t, , \tag{1}$$

where f is a nonlinear real-valued function dependent on the values of the dependent variable in some previous moments $t-1, \ldots, t-p'$, predictors $\mathbf{X}_t = (X_{1,t}, X_{2,t}, \ldots, X_{m,t})$ in the previous and/or current time moments and (ε_t) is a white noise process $(\varepsilon_t) \in N(0, \sigma^2)$.

2.2 CART method

Classification and regression tree (CART) method, known also as decision trees is a powerful data mining technique used in medicine, economy and many other areas [18, 42]. In the field of ecology, one of the first articles using this method is [43]. Today, the wide range of applications of CART makes it one of the top ten algorithms for intelligent processing of numerical and text data in computer science [44].

Generally, CART is a supervised learning algorithm capable to process very large or small datasets of arbitrary type of predictor variables $\mathbf{X}$ to regress a given dependent variable (response) Y. If the method will be applied for time series, it is necessary to include in $\mathbf{X}$ also appropriate date-time variables as predictors to keep sorting data

according to the fixed time increment t. CART can handle locally linear, nonlinear effects and interaction terms in a rule sequence and build the easily interpreted models.

The CART technique is based on a recursive type partitioning of the initial dataset into nonoverlapping multidimensional subregions (classes). The procedure is carried out in two basic steps—generating a "large tree" and then shrinking (pruning) it to avoid overfitting of the model. The structure of a CART tree consists of a root node, internal nodes and terminal nodes, each of them representing a subset of cases falling within a subregion. At any step, the cases in a current node τ could be or not be splitted into two child nodes by asking the unique question of the type of rule $X_{j,i} \leq \gamma_j$, where $X_{j,i}$ is the i-th value of the variable X_j in the current node, and γ_j is its threshold value. The CART algorithm selects the values of i and j from all possible variables X_j, $j = 1, 2, \ldots, m$ and γ_j from cases in the current node which gives the minimum error in predicting the dependent variable Y. This way, beginning from the root node, each node identifies a specific decision sequence of rules. In regression problems, the predicted value $\hat{Y}_t$ for cases classified in a node τ is simply equal to the average of their corresponding responses. The usual way to calculate the overall prediction error of the model is the least square error $L(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$. This error should be minimized at each splitting stage of the tree. The definition of a given node as a terminal one depends on a predefined control setting for a minimum number of observations in a parent and children nodes or another type of constraint [42, 44].

The method for pruning a regression tree is based on a sequential tree cut in a subtrees to simplify its structure and minimize some error-complexity measure [18, 44]. The terminal nodes of the obtained tree give the final decision about the distribution of dataset and model prediction of existing or new observations. This distribution could be considered as a form of multivariate distribution.

Figure 1 shows a simple example of a decision tree. The initial zero node root contains the entire learning dataset, with a dependent variable Y and two predictors—"minimum temperature" and "humidity." The first rule is the condition "min_temp $\leq$ 5 °C?". All observations, for which this rule is true, are classified in the left node 1, and the remaining in the right node 2. Analogically, the cases in node 2 are split in two—nodes 3 and 4 under the rule "humidity $\leq$ 60%?". The terminal nodes in Fig. 1 are 1, 3 and 4. The predicted model value $\hat{Y}_t$ for the cases in a given terminal node is the average of the cases Y_t found there. In fact, for each splitting rule, the predictor variables and their threshold values (i.e., 5 °C and 60% in Fig. 1) are chosen from all predictors used in the analysis and all their

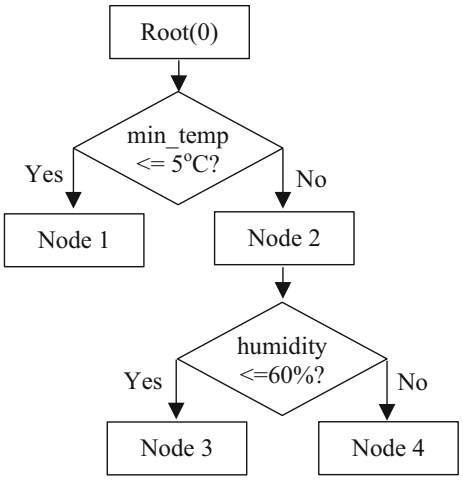

Fig. 1 An exemplary decision tree with three terminal nodes

respective values so that the overall error of the model is minimal.

In machine learning, the model performance is estimated not only for its overall accuracy, but rather than from its prediction accuracy for some known independent dataset, not used in the modeling procedure. Usually, when choosing the best pruned subtree in CART to avoid overfitting problems, a cross-validation procedure is used [18, 42, 45]. For small samples (with under 10,000 observations), the standard 10-fold cross-validation (CV) is applied [18, 42]. In the procedure of 10-fold CV, the sample is randomly divided into 10 equal nonoverlapping subsamples. Each of the subsamples is used to test the model, and the rest of the data is used as a learning sample, the process being repeated 10 times.

Main advantages of the CART method are the following: can handle equally well numerical and categorical data and combinations of these; no special requirements for the distribution and specific properties of the variables; able to detect complex nonlinear dependencies; determines the main predictors automatically; simple to understand and interpret; easy to predict new response values.

Some shortcomings of the method are: requires at least 50–100 observations and careful predictor selection; in some cases, for example in the absence of significant predictors or observations, it does not give enough accurate or stable models.

2.3 Building CART model for time series

As a regression-type technique, the choice of predictors for CART is essential to obtain an adequate model with good predictive performance. In the case of modeling a given air pollutant, consideration shall be given to the meteorological data and other pollutants, for example, precursors (if there are such and there is collected data), etc. In the

proposed method, it is recommended to use time variables as predictors. On the one hand, it describes more accurately the dependence on time, mainly of nonlinear type. Autoregressive variables from the initial time series Y are used as predictors, as well as lagged variables from other predictors, as described in Eq. (1).

Another important element of CART modeling is that according to the problem formulation in Sect. 2.1, the time series may contain different types of nonlinearities and specifics. i.e., it is not required a detailed preliminary investigation of their properties such as seasonality, stationarity, type of distribution, heteroscedasticity, etc. These properties could influence the model implicitly in the classification procedure.

The CART algorithm in the stage of generating the binary tree requires presetting limiting control settings and stopping criteria, such as minimum number of observations in a parent node, minimum number of observations in a child node, splitting method (usually least squares), type of cross-validation, maximum tree depth, etc. The result is usually numerous solutions, out of which the best model is chosen under the standard criteria, described in the next Sect. 2.4.

In our study, predictors will be included in groups, depending on the type of available data: group of lagged response variables, group of date-time variables, group of meteorological variables, group of lagged meteorological series, etc. This way, we strive to assess the impact of each group on the model. In order to more accurately identify the most important predictors of each group, pre-CART models of this group, cluster or other classification analysis, PCA analysis in the presence of multicollinearity between the predictors, etc., can be constructed and explored.

The date-time variables for a day and month of the year are defined in the following form:

$$y\_d = \text{year} + \frac{n_0 - 0.5}{365}, \tag{2}$$

$$y\_m = \text{year} + \frac{m_0 - 0.5}{12}, \tag{3}$$

where year is the initial year, n_0, m_0 are respective sequential numbers of the starting day and month from the year of the investigated time period.

The control settings will be set in a standard way, for example a minimum number of observations in parent node 10, a minimum number of observations in a child node, usually set to 5 [45, p. 306]. The latter will allow better modeling and analyzing the influence of outliers in the dataset. To avoid the potential overfitting of the models in the pruning stage, in this study we use the standard procedure with 10%-fold cross-validation [42, 45].

2.4 Model selection and evaluation

The best model is selected at each stage of involving a new group of predictors. When the control settings are specified, it is possible to obtain a single optimal solution. This does not always result in the best (in terms of the most suitable for the data) model. The reason is that model evaluation criteria are contradictory [46]. For example, if we are looking for a model with the highest coefficient of determination R^2, it will most certainly be predetermined since for input models when the number of parameters is increased, R^2 also increases. By setting more general limitations, a class of tentative models, possibly in the order of tens and hundreds of solutions, is obtained. This is not a disadvantage, but allows the researcher to select among them the most appropriate model for the given data.

When evaluating tentative best models from a given class with given predictors and control settings, we will use the widely popular goodness-of-fit criteria such as the root-mean-square error (RMSE) and the coefficient of determination R^2 defined by

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} \varepsilon_t^2}{n}}, \quad R^2 = \frac{\sum_{t=1}^{n} \left(\hat{Y}_t - \bar{Y}\right)^2}{\sum_{t=1}^{n} \left(Y_t - \bar{Y}\right)^2}, \tag{4}$$
$$\varepsilon_t = Y_t - \hat{Y}_t,$$

where $\hat{Y}_t$ is the prediction of Y_t at time moment t, $\bar{Y}$ is the sample mean, ε_t is the error.

For candidate best models, optimal model and those with the smallest RMSE and the highest R^2 within one standard error (1-SE rule) from the optimal model will be preferred. It should be noted, however, that if the objective of the analysis is prediction, different strategies for choosing the size of the tree exist [42, 45]. Practically, avoiding excessive complexity, every tree of the 1-SE set is still just as accurate and acceptable.

One of the most important steps in model diagnosis is to perform a set of tests to determine the adequacy of tentative candidates for best models. The common assumptions for validation of the regression model of the time series require that the model residuals be white noise (see Eq. 1), without serial correlation. However, the author of [8] states that "in fact, there is currently no general diagnostic statistics for nonlinear autocorrelation relationships." This is particularly true for nonlinear models for time series where the lagged series of the dependent variable and predictors are involved in. For these models, Ljung–Box [47] and other similar statistics, which are valid for the assumption of linearity, cannot be used [48]. For nonlinear time series models, it is recommended to study the autocorrelation functions (ACF) of the model residual ε_t [48].

If the CART model residuals have deviations beyond the standard limits of ACF, we recommend smoothing ε_t by an

appropriate moving average process $\hat{\varepsilon}_t = \mathrm{MA}(q)$. This will result in adjusted residual values with the error corrections $\varepsilon_t = \hat{\varepsilon}_t + e_t$. The final CART model in this case will be

$$\hat{Y}\hat{Y}_t = \hat{Y}_t + \hat{\varepsilon}_t, \quad e_t = Y_t - \hat{Y}\hat{Y}_t, \tag{5}$$

where the error e_t is assumed to be a white noise.

Finally, if the selected fitted model passes the considered tests, it can be used to make a forecast. Considering the actual use of CART models, they are used for a 1-day-ahead forecast, provided predictor values are known.

To measure overall practical forecasting accuracy of the model in this study, we use contingency tables [49], taking into account the extent to which the values exceeding the permissible limit are correctly predicted according to the established European and national standards for air quality.

2.5 ARIMA transfer function models for comparison with CART

In a large number of studies, the quality of the models is compared against models built using other statistical methods (see, e.g., [9, 10, 20]). With this approach, using the same data and selected groups of predictors, we compare the obtained best CART models against such built using the ARIMA methodology [3].

In ARIMA models, every current term of the time series Y is a linear combination of values from a fixed number of its preceding terms. When Y depends only on time t, the general form of the multiplicative ARIMA (p, d, q) model is denoted by nonnegative integer parameters p, d, q, where p is the order of an autoregressive non-seasonal process (AR); d is the order of non-seasonal trend processes (I), identified by a differencing procedure until the time series becomes stationary; q is the order of the moving average non-seasonal process (MA).

The univariate ARIMA (p, d, q) model is written in the form [3]

$$\phi_p(B)(1-B)^d \hat{Y}_t = \theta_q(B)a_t + c, \tag{6}$$

where $\phi_p(B) = \sum_{j=0}^{p} \phi_j B^j$, $\theta_q(B) = \sum_{j=0}^{q} \theta_j B^j$ are difference polynomials for autoregressive and moving average terms, respectively, B denotes the backward lag distance operator, defined as $BY_t = Y_{t-1}, \ldots, B^k Y_t = Y_{t-k}$, $a_t \sim WN(0, \sigma^2)$ is a white noise stochastic term, and c is the model constant. If the initial time series is non-stationary, it could be transformed into stationary by some degree of initial differencing, denoted by the differencing operator $(1 - B)$ in (6).

In the case, when time series Y depends on time series $\mathbf{X} = (X_1, X_2, \ldots, X_m)$, a multiple-input transfer function model can be constructed in the form [3]

$$\hat{Y}_t = \sum_{j=1}^{m} \frac{\omega_j(B)}{\delta_j(B)} B^{b_j} X_{jt} + \frac{\theta_a(B)}{\Phi_a(B)} a_t, \tag{7}$$

where $\omega_j(B), \delta_j(B), \theta_a(B), \Phi_a(B)$ are finite-order difference polynomials of B which may include AR, MA and differencing terms I, $b_j \geq 0$ are time delays.

ARIMA and transfer function methods are applicable under the assumptions of normality of the involved time series, small normally distributed residuals as a white noise, lack of autocorrelation, which could be established by using statistical tests (Ljung–Box or others), etc. In addition, in contrast to the automatic choice of the main predictors in the CART, finding the ARIMA model (p, d, q) parameters and predictor parameters is not an easy task.

3 Case studies

In this section, we provide results of the application of the proposed CART approach for modeling and forecasting pollution with particulate matter PM10 in two Bulgarian cities—Ruse and Pernik. According to a large number of scientific studies and official sources, increased concentrations of PM10 in the air cause severe and chronic respiratory diseases, cancer, allergies, etc. (e.g., [50, 51] and the literature cited therein). It has been shown that even small but constant amounts of PM10 air pollution are harmful to human health [52].

Bulgaria is among the European countries with the highest levels of PM10 pollution in urban areas and the problem persists as indicated in the latest reports by the European Environment Agency [53, 54]. In order to preserve the air quality and to comply with EU directives and standards, as well as national Bulgarian legislation, a National System for Environmental Monitoring is used, with an "air" subsystem [55, 56]. National and European regulations on monitoring, prevention, and control of air pollution and public health are governed by Directive 2008/50/EC on ambient air quality and cleaner air for Europe and standards for various pollutants [1, 2]. For particulate matter PM10, the permissible concentration values are: 24-h average of 50 μg/m^3, allowing no more than 35 exceedances within a calendar year, and up to an average of 40 μg/m^3 per year—without any exceedance. The prescribed limit values by the World Health Organization (WHO) for PM10 are: up to 20 μg/m^3 annually and up to 50 μg/m^3 per day [57].

The cities Ruse and Pernik are administrative provincial centers in Bulgaria. Ruse is located in northeastern Bulgaria in the Danube plain on the right bank of the Danube, 300 km northeast of the capital Sofia and 200 km west of the Black Sea. The city has a population of 146,000 people.

It is the most significant Bulgarian river port, serving an important part of the international trade of the country. Ruse has a temperate continental climate with very hot summers (average temperature 25 °C) and relatively cold winters (average temperature 0 °C).

Pernik is a city in western Bulgaria, located about 20 km southwest of Sofia, with a population of 80,000. It lies on both banks of the Struma River in a Pernik valley. The climate is temperate continental, and the average altitude is 750 m. Rainfall is markedly continental in nature which facilitates pollution and air self-cleaning processes. Pernik is the largest city in Bulgaria where coal is mined. Pan-European transport corridors 4 and 8 pass near the city along the Lyulin and Struma Motorways, as well as European road E871 and the railway line connecting central Europe with Greece. Among the main sources of air pollution, in addition to vehicle traffic, there are some large industrial factories.

Figure 2 shows the map of Bulgaria and the location of the two cities, marked with the star symbol.

3.1 Data

This study uses official measurements of PM10 for Ruse and Pernik, published as graphics in the annual regional reports for the two cities [58, 59], as well as data on atmospheric and meteorological variables, collected from [60, 61]. The time series analysis is based on the measured average daily concentrations of the PM10 air pollutant in Ruse over a period of 6 years, from January 1, 2011, to December 31, 2016 and for Pernik over a period of 5 years, from January 1, 2010 to December 31, 2014. All PM10 data are used in all built models, except those for the last day of the period (December 31, 2016, for Ruse and December 31, 2014 for Pernik) used for forecasting 1 day ahead. In a realistic situation, the forecasting procedure could also use the forecasts for meteorological variables. For Bulgaria, the weather forecasts are provided for many cities, including Ruse and Pernik by the European system ALADIN [62].

The variables and groups of variables used during the process of modeling and forecasting are given in Table 1.

In order to account for the periodic nature of the variable *wind_direction*, it is transformed into a new variable WDI using the expression

$$\mathrm{WDI} = 1 + \sin(wind\_direction + \pi/(k-1)), \tag{8}$$

where k is the number of different wind directions. In our case $k = 16$.

The descriptive statistics of the initial data for Ruse and Pernik are given in Tables 2 and 3, respectively. The missing measured values of PM10 are 1% for Ruse and 9.1% for Pernik. In subsequent analyses, missing values are substituted by linear interpolation. For Ruse in Table 2, the mean value is 43.18 μg/m^3 over the entire 6-year period and the maximum daily PM10 concentration is 214 μg/m^3.

Fig. 2 Map of Bulgaria with the location of the cities of Ruse (in the north) and Pernik (on the west)

Table 1 Variables used in CART models of PM10 concentrations in Ruse and Pernik, Bulgaria

Type	Variable	Description
Dependent	*Y_R* (μg/m^3)	Daily PM10 concentration for the city of Ruse
Dependent	*Y_P* (μg/m^3)	Daily PM10 concentration for the city of Pernik
Predictor group {1}	*Y_R*⟨1⟩ (μg/m^3), *Y_R*⟨2⟩ (μg/m^3)	Lagged variables of *Y_R* in the models of Ruse
	Y_P⟨1⟩ (μg/m^3), *Y_P*⟨2⟩ (μg/m^3)	Lagged variables of *Y_P* in the models of Pernik
Predictor group {2} (date-time variables)	*y_d*	Day of the year (2)
	y_m	Month of the year (3)
Predictor group {3} (meteorological variables)	*min_temp* (°C)	Minimum daily air temperature
	max_temp (°C)	Maximum daily air temperature
	wind_sp (°C)	Wind speed
	wind_direction, rad (WDI)	Wind direction
	precip (%)	Precipitation
	hum (%)	Relative humidity
	press (mb)	Air pressure
	cl_cov (%)	Cloud cover
Predictor group {4} (lagged predictors)	*min_temp*⟨1⟩, *max_temp*⟨1⟩, *precip*⟨1⟩, etc.	Lagged meteorological variables from group {3} for one and two past days

Table 2 Descriptive statistics of initial data for the city of Ruse

Statistic	*Y_R* (μg/m^3)	*min_temp* (°C)	*max_temp* (°C)	*wind_sp* (m/s)	*WDI*	*precip* (%)	*hum* (%)	*press* (mb)	*cl_cov* (%)
N	2173	2190	2190	2190	2190	2190	2190	2190	2190
N missing	17	0	0	0	0	0	0	0	0
Mean	43.18	11.74	18.98	3.24	1.03	1.71	0.68	1017	0.31
Median	36.4	12	20	2.7	1.18	0	0.68	1016	0.23
SD	26.18	10.19	11.37	1.59	0.86	4.18	0.16	7.47	0.26
Minimum	4.2	− 20	− 13	0.4	0.17	0	0.26	989	0
Maximum	214.1	32	42	12.1	1.98	39	0.99	1041	1

Table 3 for Pernik shows a high mean value of 64.20 μg/m^3 for PM10 and a maximum average daily concentration of 347.5 μg/m^3 over the examined 5-year period. This shows that this pollutant is very problematic, especially for the town of Pernik. From Tables 2 and 3, a large range between maximum and minimum temperature values both with relatively weak winds are observed. These weather conditions favor the formation of high concentrations of PM10.

Sequence charts of the PM10 time series for both cities are given in Fig. 3 and 4, where the horizontal line is the prescribed European and national upper limit for the average daily concentration of PM10 of 50 μg/m^3. There are numerous exceedances over the entire periods. More specifically, for Ruse, the measured exceedances are 571 (or 26.3%), and for Pernik, these are 786 (47.4%). Relatively high exceedances are seen during winter, and the lowest are in summer. Along with meteorological factors, 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 [58, 59], 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 factors.

The type of PM10 time series observed in Fig. 3 and 4 leads to the conclusion that there are no trends and the

Table 3 Descriptive statistics of initial data for the city of Pernik

Statistic	*Y_P* (μg/m^3)	*min_temp* (°C)	*max_temp* (°C)	*wind_sp* (m/s)	*WDI*	*precip* (%)	*hum* (%)	*press* (mb)	*cl_cov* (%)
N	1659	1825	1825	1825	1825	1825	1825	1825	1825
N missing	166	0	0	0	0	0	0	0	0
Mean	64.20	8.87	16.95	2.10	0.96	1.67	0.73	1016	0.26
Median	48.30	9	17	1.79	0.88	0.20	0.73	1016	0.2
SD	51.17	9.7	10.08	0.76	0.74	4	0.13	6.72	0.21
Min	0.44	− 20	− 10	0.45	0.02	0	0.33	991	0
Max	347.5	28	38	5.81	1.98	52.9	1	1038	1

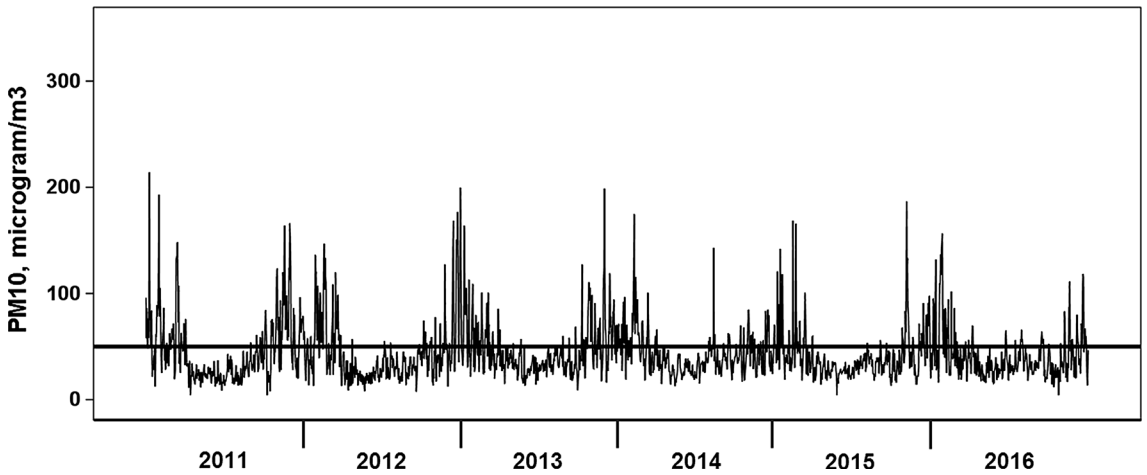

Fig. 3 Sequence plot of the initial time series of PM10 concentration for the town of Ruse, on daily basis. The horizontal line of 50 μg/m^3 indicates the permissible daily limit value according to national and European standards

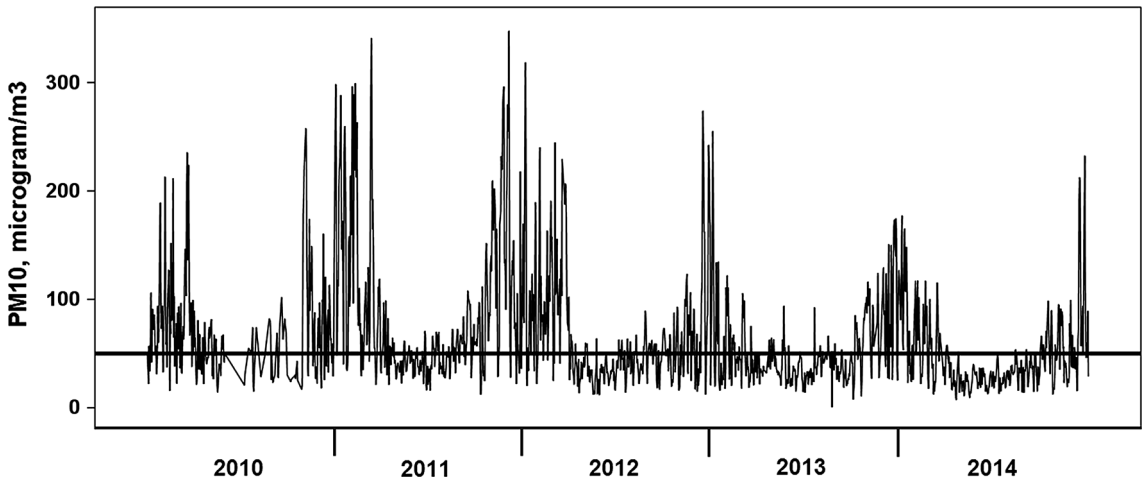

Fig. 4 Sequence plot of the initial time series of PM10 concentration for the town of Pernik, on daily basis. The horizontal line of 50 μg/m^3 indicates the permissible daily limit value according to national and European standards

series are stationary. This was confirmed by the unit root test and the ADF test that showed values 1.12×10^{-10} for *Y_R* and 2.34×10^{-10} for *Y_P*, respectively. Annual seasonality is also observed from Figs. 3 and 4.

Overall, the two time series of PM10 are quite similar in type, with the cities being 275 km apart.

3.2 CART modeling results

Following the common modeling approach described in Sect. 2, numerous models were built and analyzed. The main statistics of the best CART models obtained for Ruse and Pernik are given in Tables 4 and 5, respectively. The

Table 4 Summary of the performance of the obtained tentative CART models for Ruse

Model	Predictor groups	RMSE learn	R^2 learn	Number of the TNodes	Main predictors in the model, sorted by their relative importance
R1	{1, 2}	17.337	0.562	8	*Y_R*⟨1⟩, *Y_R*⟨2⟩, *y_d*, *y_m*
RM1	{1, 2}	16.806	0.588	14	*Y_R*⟨1⟩, *Y_R*⟨2⟩, *y_d*, *y_m*
R2	{1, 2, 3}	15.048	0.670	19	*Y_R*⟨1⟩, *Y_R*⟨2⟩, *min_temp*, *max_temp*, *wind_sp*, *hum*, *y_d*, *press*, *cl_cov*, *precip*, *WDI*, *y_m*
RM2	{1, 2, 3}	13.117	0.749	45	*Y_R*⟨1⟩, *Y_R*⟨2⟩, *min_temp*, *max_temp*, *wind_sp*, *hum*, *y_d*, *press*, *cl_cov*, *precip*, *WDI*, *y_m*
*RS2	S2 from {1, 2, 3}	12.325	0.779	85	*Y_R*⟨1⟩, *Y_R*⟨2⟩, *min_temp*, *max_temp*, *wind_sp*, *hum*
R3	{1, 2, 3, 4}	13.600	0.730	34	*Y_R*⟨1⟩, *Y_R*⟨2⟩, *min_temp*⟨1⟩, *max_temp*⟨1⟩, *min_temp*, *wind_sp*, *hum*⟨1⟩, *precip*, *max_temp*, …
RM3	{1, 2, 3, 4}	12.525	0.771	58	*Y_R*⟨1⟩, *Y_R*⟨2⟩, *min_temp*⟨1⟩, *max_temp*⟨1⟩, *min_temp*, *wind_sp*, *hum*⟨1⟩, *max_temp*, *precip*
RS3	S3 from {1, 2, 3, 4}	12.527	0.771	77	*Y_R*⟨1⟩, *Y_R*⟨2⟩, *min_temp*⟨1⟩, *min_temp*, *max_temp*⟨1⟩, *wind_sp*, *hum*⟨1⟩
RLag	S2 from {1, 2, 4}	14.331	0.700	33	*Y_R*⟨1⟩, *Y_R*⟨2⟩, *min_temp*⟨1⟩, *max_temp*⟨1⟩, *wind_sp*⟨1⟩, *hum*⟨1⟩, *y_d*

Table 5 Summary of the performance of the obtained CART models for Pernik

Model	Predictor groups	RMSE learn	R^2 learn	Number of the TNodes	Main predictors in the model, sorted by their relative importance
P1	{1, 2}	29.014	0.667	10	*Y_P*⟨1⟩, *Y_P*⟨2⟩, *y_d*, *y_m*
PM1	{1, 2}	26.196	0.729	29	*Y_P*⟨1⟩, *Y_P*⟨2⟩, *y_d*, *y_m*
P2	{1, 2, 3}	23.929	0.774	22	*Y_P*⟨1⟩, *Y_P*⟨2⟩, *press*, *min_temp*, *precip*, *max_temp*, *hum*, …
PM2	{1, 2, 3}	19.177	0.855	74	*Y_P*⟨1⟩, *Y_P*⟨2⟩, *press*, *min_temp*, *precip*, *hum*, *max_temp*, *y_d*, …
PS2	S2 from {1, 2, 3}	20.252	0.838	69	*Y_P*⟨1⟩, *Y_P*⟨2⟩, *press*, *min_temp*, *precip*, *max_temp*, *hum*
P3	{1, 2, 3, 4}	22.143	0.806	30	*Y_P*⟨1⟩, *Y_P*⟨2⟩, *min_temp*⟨1⟩, *cl_cov*, *precip*, *min_temp*, *hum*, *press*, *max_temp*, …
*PM3	{1, 2, 3, 4}	15.573	0.904	289	*Y_P*⟨1⟩, *Y_P*⟨2⟩, *min_temp*⟨1⟩, *cl_cov*, *min_temp*, *precip*, *hum*, *press*, *max_temp*, *max_temp*⟨1⟩, *y_d*, …
PS3	S3 from {1, 2, 3, 4}	17.983	0.872	130	*Y_P*⟨1⟩, *Y_P*⟨2⟩, *min_temp*⟨1⟩, *press*, *cl_cov*, *precip*
PLag	S4 from {4}	19.962	0.842	105	*Y_P*⟨1⟩, *min_temp*⟨1⟩, *press*⟨1⟩, *max_temp*⟨1⟩, *y_d*, *cl_cov*

models marked with R and P are optimal for the selected predictors for Ruse and Pernik. The RM and PM models that follow them have maximum values of the corresponding coefficients of determination, selected from all accurate models within one standard error of minimum (1-SE rule) [18, 42]. The S symbol is used for models RS and PS that are built only with the main predictors that have more than 10 units of relative importance in the model. The last models in the two tables, marked with RLag and PLag, are the best CART models constructed using predictor groups {1, 2} and lagged predictors from group {4} in accordance with the specific purpose (i), stated in the introduction.

Table 4 for Ruse models shows that RMSE takes values from 17.3 to 12.3 and the coefficient of determination R^2 varies from 0.56 to 0.78. The selected best model is denoted by an asterisk (see model *RS2 in line 5 of Table 4), with a minimum RMSE = 12.325 and a maximum R^2 = 0.779. Figure 5a shows the resulting error curve against the tree size for models with predictors from group S for Ruse. It is observed the almost equal accuracy of the 1-SE models.

Table 5 represents summary for the selected tentative CART models for the city of Pernik. RMSE varies from 29 to 15.6, and the coefficients of determination R^2 change from 0.67 to 0.90. The statistical indexes of the models show that the best features are obtained for model *PM3 with a minimum RMSE = 15.57 and a maximum value of $R^2 = 0.904$. Figure 5b shows the corresponding error curve against the number of the terminal nodes for models with predictors from groups {1, 2, 3, 4} for the city of Pernik.

The relative importance of the predictors of the best PM10 model *RS2 for Ruse is presented on the left-hand side of Table 6. The information about the best selected model *PM3 for Pernik is given on the right-hand side of Table 6.

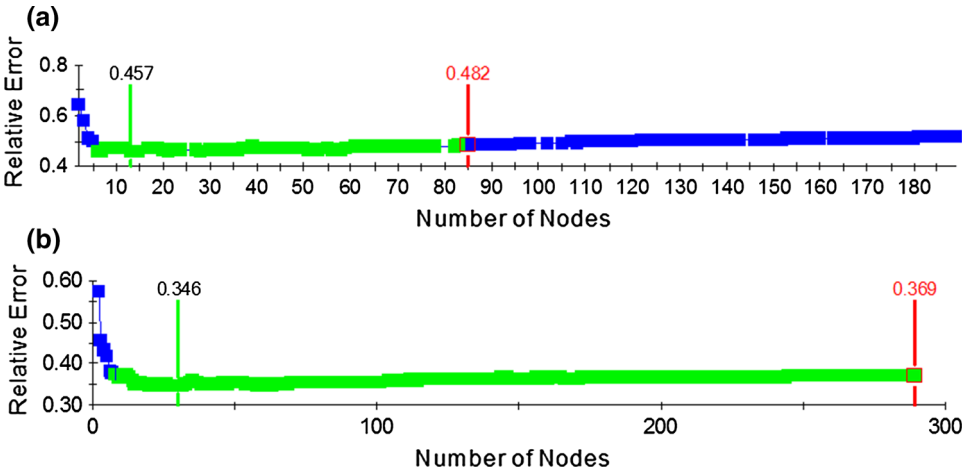

Fig. 5 Relative error curves against the number of the terminal nodes, where 1-SE models are in green color: **a** with predictors from S2-set for Ruse; **b** with predictors from groups {1, 2, 3, 4} for Pernik. The left vertical line indicates the optimal model and the right vertical line the selected model

Table 6 Relative variable importance of the predictors in models *RS2 and *PM3

Model *RS2 for Ruse		Model *PM3 for Pernik[a]	
Variable	Score	Variable	Score
Y_R⟨1⟩	100	*Y_P*⟨1⟩	100
Y_R⟨2⟩	48.0	*Y_P*⟨2⟩	49.0
min_temp	28.6	*min_temp*⟨1⟩	16.5
max_temp	20.3	*cl_cov*	12.4
wind_sp	20.1	*min_temp*	11.8
hum	18.9	*precip*	11.8
		max_temp⟨1⟩	10.4
		hum	10.4
		press	10.0
		max_temp	9.9
		y_d	6.6

[a]The variables in the *PM3 model of least relative importance are omitted

The diagnosis of the residuals of the two selected best models by computing residual ACF (see [48]) showed that they do not have autocorrelation. For the best model *RS2, an error correction of residuals was performed using formula (6) with a moving average process MA (7). The resulting model is denoted by $\hat{Y}\hat{Y}\_R$. Analogous error correction was also obtained for the *PM3 model with a MA (28) process and the fitted model was denoted by $\hat{Y}\hat{Y}\_P$. Figure 6a, b shows the residual ACF of the models before and after the corrections. From Fig. 6c, d, it can be seen that the ACF of the residuals e_t of the final models are small enough and the autocorrelation is removed. Heteroscedasticity in the models was not observed. Therefore, we can assume that the selected models are consistent with the studied data.

3.3 Assessment of the predictive and forecasting qualities of the CART models considering EU standards for PM10

In order to achieve a practical assessment of model results, we will compare their fit performance to the measured PM10 concentrations and will also apply these to obtain a 1-day-ahead forecast. The values predicted by the models are compared to the observed values of PM10 against the permissible threshold set by the European union of 50 μg/m^3 per day. Figure 7a shows observed values compared to the ones predicted by the selected best model for Ruse $\hat{Y}\hat{Y}\_R$. Figure 7b analogically shows the results for the model $\hat{Y}\hat{Y}\_P$ for the city of Pernik. Horizontal and vertical lines correspond to the permissible average daily threshold of 50 μg/m^3. The four rectangular areas resulting from the intersection of these straight lines contain the predicted points, which are correctly or incorrectly predicted according to the threshold value. Correctly predicted values over 50 μg/m^3 are located in the right upper area, and those correctly predicted below 50 μg/m^3 are located in the lower left area.

The correctly predicted results, taking into account data used in the analyses and actual measured values, are given in 2 × 2 contingency tables—Table 7 and 8, respectively, for the two cities (for the contingency tables, see for example [40]). Table 7 shows that for Ruse, the number of predictions under the permissible 50 μg/m^3 is 1520, which represents 95% of the observed PM10 values under 50 μg/m^3. The number of correct predictions over 50 μg/m^3 is 421 or 74% of the number of the measured exceedances. The total number of correctly classified cases in the two groups is 89%. The incorrectly classified cases are 81 for < 50 μg/m^3 and 150 for ≥ 50 μg/m^3. We can conclude that the obtained model *RS2 for Ruse shows excellent predictive properties. The model is applied to forecast pollution in Ruse for 1 day ahead, i.e., for December 31,

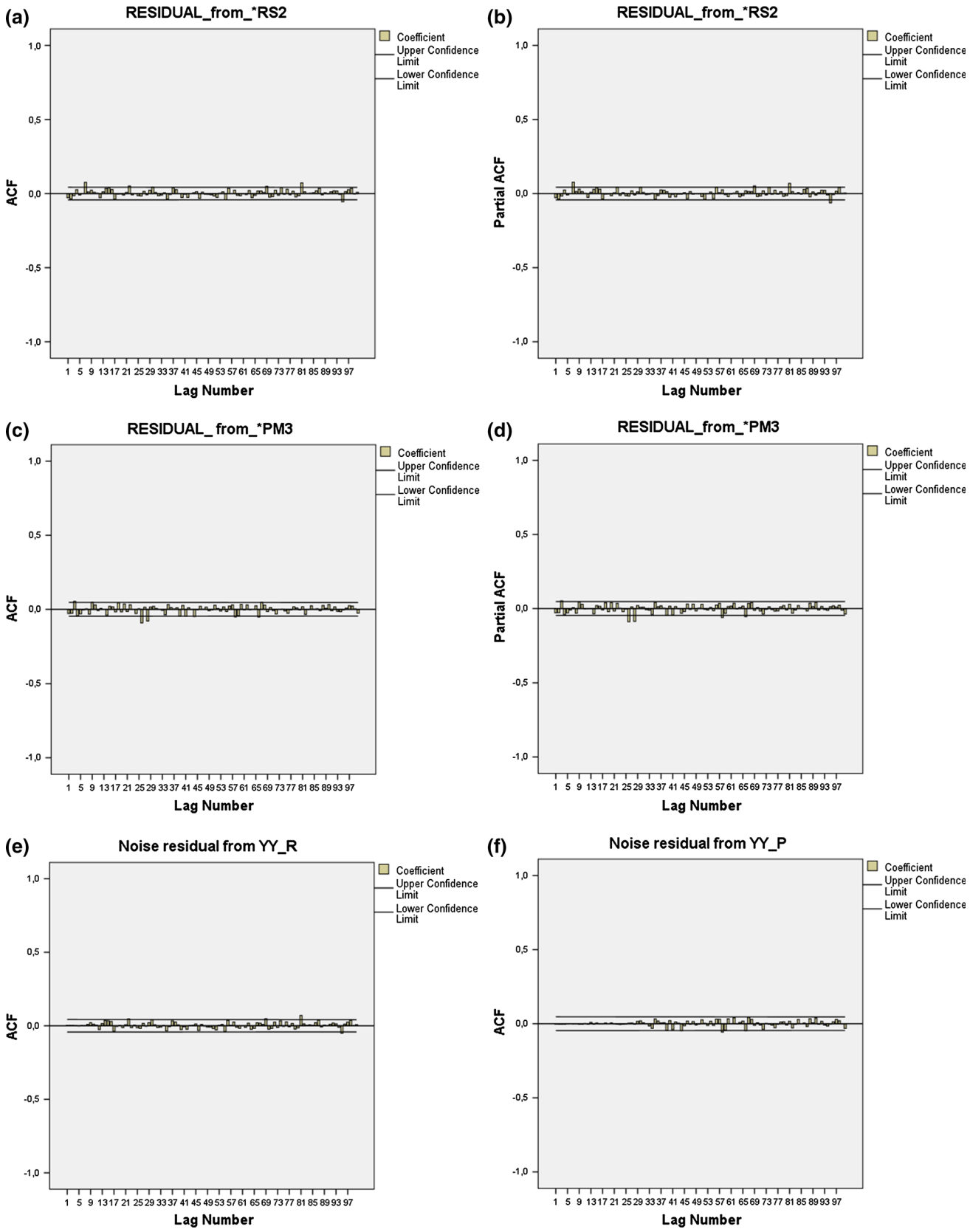

Fig. 6 **a**, **b** Residual autocorrelation function (ACF) and partial ACF (PACF) of the obtained best model *RS2; **c**, **d** residual ACF and PACF of the obtained best model *PM3; **e**, **f** ACF of error corrected models $\hat{Y}\hat{Y}\_R$ and $\hat{Y}\hat{Y}\_P$, respectively

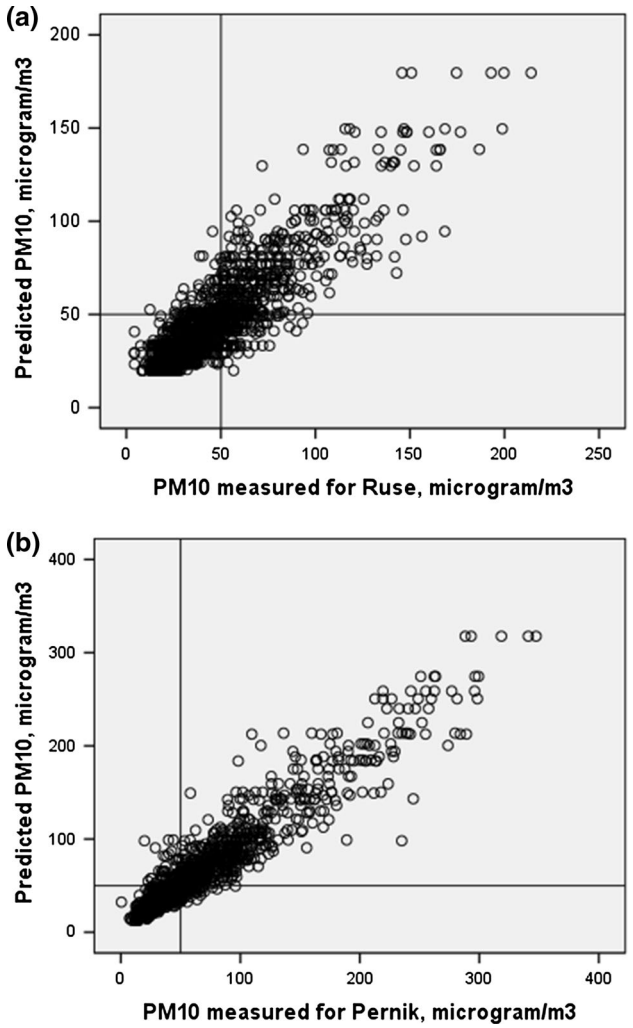

Fig. 7 Predicted PM10 concentrations versus measured data using the best models: **a** $\hat{Y}\hat{Y}\_R$ for Ruse; **b** $\hat{Y}\hat{Y}\_P$ for Pernik. The horizontal and vertical lines depict the European upper permissible daily average PM10 pollution of 50 μg/m^3

Table 7 Contingency table for model $\hat{Y}\hat{Y}\_R$ considering only measured PM10 data for Ruse

Model: $\hat{Y}\hat{Y}\_R$	Forecast			
	< 50	≥ 50	Total	$\%O$
Observation				
< 50	1520	81	1601	95
≥ 50	150	421	571	74
Total	1670	502	2172	89
$\%P$	91	84		

2016. Figure 8a illustrates the last measured and predicted PM10 values for 30 days and the forecast.

Table 8 Contingency table for model $\hat{Y}\hat{Y}\_P$ considering only measured PM10 data for Pernik

Model: $\hat{Y}\hat{Y}\_P$	Forecast			
	< 50	≥ 50	Total	$\%O$
Observation				
< 50	789	83	872	90
≥ 50	71	715	786	91
Total	860	798	1658	91
$\%P$	92	90		

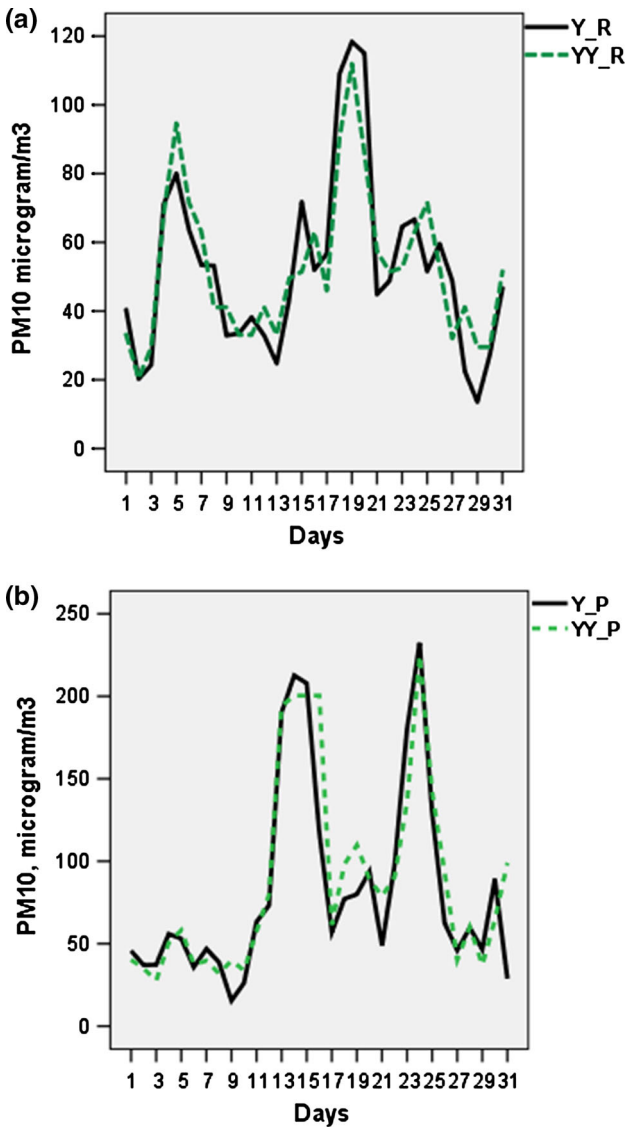

Fig. 8 Comparison of the last 31 actual data (solid line) against predictions from models (dashed line), where the last point shows the forecast: **a** from model $\hat{Y}\hat{Y}\_R$ for Ruse; **b** from model $\hat{Y}\hat{Y}\_P$ for Pernik

Table 9 Summary of the performance of the obtained ARIMA (1, 0, 5) transfer function models for Ruse and Pernik

Model	Predictor group	RMSE	R^2	Ljung–Box	Significant predictors with (p, d, q) parameters
ARM2	{1, 2, 3}	15.628	0.645	0.463	*Y_R*⟨1⟩; *min_temp*(0, 0, 0); *max_temp*(0, 0, 0)-delay 2; *wind_sp*(0, 0, 2); *hum*(0, 0, 0)-delay 1; *press*(0, 0, 0)-delay 1; *cl_cov*(0, 0, 0); *precip*(0, 0, 0)
A*RS2	S2 from {1, 2, 3}	16.436	0.608	0.073	*Y_R*⟨1⟩, *min_temp*(0, 0, 0)-delay 1; *max_temp*(0, 0, 0)-delay 1; *wind_sp*(0, 0, 0); *hum*(0, 0, 0)-delay 1
ARM3	{1, 2, 3, 4}	15.962	0.632	0.473	*Y_R*⟨1⟩, all variables form {3, 4} with (0, 0, 0)
ARLag	{1, 2, 4}	16.444	0.610	0.797	*Y_R*⟨1⟩, *min_temp*⟨1⟩(0, 0, 2)-delay 1; *hum*⟨1⟩(0, 0, 0); *wind_sp*⟨1⟩(0, 0, 1); *WDI*(0, 0, 2)-delay 2; *press*⟨1⟩(0, 0, 6)
ARLagS	S2 from {1, 2, 4}	16.929	0.584	0.677	*Y_R*⟨1⟩, *min_temp*⟨1⟩(0, 0, 2)-delay 1; *hum*⟨1⟩(0, 0, 0); *wind_sp*⟨1⟩(0, 0, 1)
APM2	{1, 2, 3}	25.494	0.746	0.951	*Y_P*⟨1⟩, *min_temp*(0, 0, 7); *max_temp*(0, 0, 0); *wind_sp*(0, 0, 1); *hum*(0, 0, 0); *press*(0, 0, 0)-delay 1; *cl_cov*(0, 0, 1); *precip*(0, 0, 0), *WDI*(0, 0, 0)
A*PM3	{1, 2, 3, 4}	25.162	0.752	0.957	*Y_P*⟨1⟩, all variables from{3, 4} with (0, 0, 0)
APLag	{1, 2, 4}	28.910	0.672	0.071	*Y_P*⟨1⟩, all variables from {4} with (0, 0, 0)

Analogically, in Table 8 for the best model *PM3 for Pernik, we obtain very similar results with a total of 91% correctly classified values. Figure 8b illustrates actual and predicted PM10 concentrations for December 2014, as well as the forecasted value for the last day December 31, 2014.

3.4 Building ARIMA transfer function models and comparison with the CART results

To compare the performance of the best CART models with ARIMA models and check the forecast ability according (i), (ii) and (iii) conditions, stated in the introduction, we apply ARIMA method with the corresponding groups of predictor variables from Tables 4 and 5. The correspondence of ARIMA with CART models is denoted by the first symbol A.

ARIMA analysis is performed following the usual procedure of several steps [3]:

1. Identification of model parameters
2. Build and analyze ARIMA transfer function models for different groups of predictors.
3. Evaluate model accuracy and adequacy of the models regarding goodness of fit criteria (4), error residuals, assessing whether the residuals exhibit autocorrelation using Ljung–Box test. Choose the best models.
4. Apply the selected best models to predict PM10 concentrations.

Using autocorrelation and partial autocorrelation functions (ACF, PACF) for both initial time series *Y_R* and *Y_P,* it was obtained very similar behavior, without trends and seasonality. The model type for both Ruse and Pernik was identified as ARIMA (1, 0, 5). This means that in all models *Y_R*⟨1⟩ (respectively *Y_P*⟨1⟩) are included in group {1} and the two-lagged variables *Y_R*⟨2⟩, *Y_P*⟨2⟩ are not.

To improve distribution and stabilize the variability, the *Y_R* and *Y_P* were initially transformed by natural logarithm. There were also transformed by natural logarithm the following predictors: *wind_sp*, *hum*, and *press* for Ruse, and *wind_sp*, *WDI*, *hum* and *press* for Pernik. The best ARIMA transfer function models with nonlagged, lagged and mixed type predictors are shown in Table 9. All model parameters are obtained at significance level $\alpha = 0.01$.

4 Discussion with conclusion

In Sect. 2, a common CART approach is proposed that addresses the general nature of nonlinear dependence and general conditions aimed at modeling air pollutants and similar types of time series in environmental sciences. Key aspects are: inclusion of predictor groups from autoregressive variables of the dependent variable, independent predictors, autoregressive predictor variables, specific time-counting variables, and combinations of predictor groups. Assuming large samples of data, different aspects of CART models are explored, including construction of stable models by stepwise predictor groups modeling, correcting CART model errors through moving average (MA) process, improving model forecasting procedure according to the type of predictors (lagged and/or unlagged), assessment against European and national health standards, and others. Also, the use of maximum models in the 1-SE rule area, in practice, allows for more detailed study of outliers and other local conditions following the rules of the CART tree.

Applying the proposed common approach empirical PM10 data for two Bulgarian cities—Ruse and Pernik, are investigated. The obtained models have similar performance. This could be explained by the stochastic nature of PM10 data, which implicitly contains the influence of similar factors for the two cities. It can be noted that for both cities the weather conditions are combined with major sources of PM10 pollution such as the use of solid fuels for domestic and industrial purposes in winter and, to a lesser extent, road traffic, according to regional reports [58, 59].

A very important result of CART models is the ability to clearly determine the relative strength of importance of each predictor in the model. Table 6 shows that the lagged variables of *Y_R* and *Y_P* for PM10 have the highest influence on the models, with observations from the two previous days of the current one. In the best model for Ruse (model *RS2), the corresponding predictors are $Y\_R\langle 1\rangle$ and $Y\_R\langle 2\rangle$ with a relative importance of 100% and 48%, respectively. For Pernik in the *PM3 model, the respective variables $Y\_P\langle 1\rangle$ and $Y\_P\langle 2\rangle$ have a relative importance of 100% and 49%. These autoregressive terms contain implicitly the influence of all factors for formation, accumulation or dispersion of particulate matter—sources of pollution, weather conditions, levels of PM10 pollution precursors such as SO2, CO and NO, climate and geographic location, etc.

Other similar factors in the constructed CART models are meteorological data, including minimum and maximum average daily temperatures, and humidity, which are more favorable for formation and accumulation of PM10. For Ruse, the *RS2 model is also influenced by the wind speed, according to the flat relief of the town. In the case of Pernik, located in the valley, the weaker winds (see Table 3) are not affected the model *PM3, and the weather conditions are more stable during periods of 1 and 2 days. The latter explains the effect of the minimum air temperature of both the previous and the current day, cloud cover and precipitation.

Comparing the increase of R^2 from Tables 4 and 5, we can conclude that the group of meteorological factors improves the quality of the models within 20–25%. Also, the role of the predictor group {4} with lagged meteorological variables does not significantly differ from the importance of predictor group {3}. Particularly, models RLag and PLag have lower R^2 of about 5–8% compared to RM2, *RS2 and PM2. The same relation is observed in the *PM3 model with mixed type predictors. Hence, the choice of different types of predictors under schemes (i), (ii) or (iii) will have relatively small impact in the CART models from 5 to 20%, with the best results coming from the mixed groups of predictors {1, 2, 3, 4}. It can be expected that when using forecast of meteorological predictors for time $t + 1$ (case (i)) to forecast Y_{t+1} the error will not be significantly different from the same forecast by only lagged variables (ii). More exactly, in the best model *PM3 the influence of weather forecasts (meteorological conditions for a future period) will have almost the same relative weights with corresponding lagged predictors. For the minimum temperature, these weights in Table 6 are, respectively, 11.8 and 16.5, and for the maximum temperature they are 10.4 and 9.9. Consequently, the use of CART models with mixed predictors of group type {1, 2, 3, 4} can be recommended to minimize the errors of future PM10 predictions.

The CART results from the contingency tables (see Tables 7, 8) are also similar for the two cities. More exactly, 89% of the actual PM10 measurements for Ruse are correctly classified against the limit of 50 $\mu g/m^3$. 91% of the data is correctly classified for Pernik.

Finally, comparison of CART and ARIMA methods was conducted. The results from the goodness-of-fit criterion R^2 from Table 9 and Tables 4 and 5 lead to the following conclusions: CART models for Rouse with a set of predictors {1, 2, 3} have higher R^2 values than ARIMA models within 10–17% and those for Pernik 11%, respectively. The R^2 differences between CART and ARIMA models for Ruse with the group with lagged predictors {1, 2, 4} are 9% (for RLag and ARLag) and 17% (for Plag and APLag) in favor of CART. For group {1, 2, 3, 4}, the corresponding R^2 differ with 14% and 15%. Also the accuracy evaluated by the RMSE of all obtained CART models is better related to the RMSE of the ARIMA models. In general, the results of the CART models outperform those of ARIMA method.

From the conducted empirical studies, we can conclude that the proposed CART methodology produces very good results and demonstrates high statistical performance and fitting with the investigated data. The obtained models could be successfully applied for the analysis, forecasting and control of PM10 pollution levels in urban environments depending on meteorological, atmospheric and other conditions.

Acknowledgements This work was supported by the Grant No. BG05M2OP001-1.001-0003, financed by the Science and Education for Smart Growth Operational Program (2014–2020), co-financed by the European Union through the European structural and Investment funds. We want to express our gratitude to the independent reviewers for the valuable advice and feedback, which helped improve the scientific value of this study.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. 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 L 152/1. https://eur-lex.europa.eu/eli/dir/2008/50/oj. Accessed 15 July 2019
2. Air Quality Standards (2015) European Commission. Environment. http://ec.europa.eu/environment/air/quality/standards.htm. Accessed 15 July 2019
3. Box GEP, Jenkins GM, Reinsel GS (1994) Time series analysis, forecasting and control, 3rd edn. Prentice-Hall Inc., Upper Saddle River
4. Liu PWG (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. https://doi.org/10.1016/j.atmosenv.2009.01.055
5. Pohoata A, Lungu E (2017) A complex analysis employing ARIMA model and statistical methods on air pollutants recorded in Ploiesti, Romania. Rev Chim 68(4):818–823
6. Stoimenova M (2016) Stochastic modeling of problematic air pollution with particulate matter in the city of Pernik, Bulgaria. Ecol Balk 8(2):33–41
7. Zheleva I, Veleva E, Filipova M (2017) Analysis and modeling of daily air pollutants in the city of Ruse. Bulgaria. AIP Conf Proc 1895:030007. https://doi.org/10.1063/1.5007366
8. Zhang PG (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50:159–175. https://doi.org/10.1016/S0925-2312(01)00702-0
9. Lee NU, Shim JS, Ju YW, Park SC (2017) Design and implementation of the SARIMA–SVM time series analysis algorithm for the improvement of atmospheric environment forecast accuracy. Soft Comput. https://doi.org/10.1007/s00500-017-2825-y
10. Nieto PJG, Lasheras FS, García-Gonzalo E, de Cos Juez FJ (2018) PM10 concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: a case study. Sci Total Environ 621:753–761. https://doi.org/10.1016/j.scitotenv.2017.11.291
11. Zhang H, Zhang S, Wang P, Qin Y, Wang H (2017) Forecasting of particulate matter time series using wavelet analysis and wavelet-ARMA/ARIMA model in Taiyuan, China. J Air Waste Manag Assoc 67(7):776–788. https://doi.org/10.1080/10962247.2017.1292968
12. Biancofiore F, Busilacchio M, Verdecchia M, Tomassetti B, Aruffo E, Bianco S, Di Tommaso S, Colangeli C, Rosatelli G, Di Carlo P (2017) Recursive neural network model for analysis and forecast of PM10 and PM2.5. Atmos Pollut Res 8(4):652–659. https://doi.org/10.1016/j.apr.2016.12.014
13. Franceschi F, Cobo M, Figueredo M (2018) Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá, Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering. Atmos Pollut Res 9(5):912–922. https://doi.org/10.1016/j.apr.2018.02.006
14. Bougoudis I, Demertzis K, Iliadis L (2016) HISYCOL a hybrid computational intelligence system for combined machine learning: the case of air pollution modelling in Athens. Neural Comput Appl 27(5):1191–1206. https://doi.org/10.1007/s00521-015-1927-7
15. Abderrahim H, Chellali MR, Hamou A (2016) Forecasting PM10 in Algiers: efficacy of multilayer perceptron networks. Environ Sci Pollut Res 23(2):1634–1641. https://doi.org/10.1007/s11356-015-5406-6
16. Prakash A, Kumar U, Kumar K, Jain V (2011) A wavelet-based neural network model to predict ambient air pollutants' concentration. Environ Model Assess 16(5):503–517. https://doi.org/10.1007/s10666-011-9270-6
17. Morgan JN, Sonquist JA (1963) Problems in an analysis of survey data and a proposal. J Am Stat Assoc 58:415–434
18. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth Advanced Books and Software, Belmont
19. Burrows WR, Benjamin M, Beauchamp S, Lord ER, McCollor D, Thomson B (1995) CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. J Appl Meteorol 34:1848–1862. https://doi.org/10.1175/1520-0450(1995)034<1848:CDTSAA>2.0.CO;2
20. Slini T, Kaprara A, Karatzas K, Moussiopoulos N (2006) PM10 forecasting for Thessaloniki, Greece. Environ Model Softw 21(4):559–565. https://doi.org/10.1016/j.envsoft.2004.06.011
21. Zickus M, Greig AJ, Niranjan M (2002) Comparison of four machine learning methods for predicting PM10 concentrations in Helsinki, Finland. Water Air Soil Pollut Focus 2:717–729. https://doi.org/10.1023/A:1021321820639
22. Choi W, Paulson SE, Casmassi J, Winer AM (2013) Evaluating meteorological comparability in air quality studies: classification and regression trees for primary pollutants in California's South Coast Air Basin. Atmos Environ 64:150–159. https://doi.org/10.1016/j.atmosenv.2012.09.049
23. Sayegh A, Tate JE, Ropkins K (2016) Understanding how roadside concentrations of NOx are influenced by the background levels, traffic density, and meteorological conditions using Boosted Regression Trees. Atmos Environ 127:163–175. https://doi.org/10.1016/j.atmosenv.2015.12.024
24. Stoimenova M, Voynikova D, Ivanov A, Gocheva-Ilieva S, Iliev I (2017) Regression trees modeling and forecasting of PM10 air pollution in urban areas. AIP Conf Proc 1895:030005. https://doi.org/10.1063/1.5007364
25. Lewis PAW, Stevens JG (1991) Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). J Am Stat Assoc 86(416):864–877. https://doi.org/10.1080/01621459.1991.10475126
26. Weber G-W, Batmaz I, Köksal G, Taylan P, Yerlikaya-Özkurt F (2012) CMARS: a new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization. Inverse Probl Sci Eng 20(3):371–400. https://doi.org/10.1080/17415977.2011.624770
27. Özmen A, Weber G-W, Batmaz I (2010) The new robust CMARS (RCMARS) method. In: Kasımbeyli R, Dinçer C, Özpeynirci S, Sakalauskas L (eds) 24th mini EURO conference on continuous optimization and information-based technologies in the financial sector, MEC EurOPT 2010, pp 362–368
28. Özmen A, Weber GW (2012) Robust conic generalized partial linear models using RCMARS method—a robustification of CGPLM. AIP Conf Proc 1499:337–343. https://doi.org/10.1063/1.4769011
29. Özmen A, Weber G-W (2014) RMARS: Robustification of multivariate adaptive regression spline under polyhedral uncertainty. J Comput Appl Math 259(Part B):914–924. https://doi.org/10.1016/j.cam.2013.09.055
30. Özmen A, Batmaz İ, Weber G-W (2014) Precipitation modeling by polyhedral RCMARS and comparison with MARS and CMARS. Environ Model Assess 19(5):425–435. https://doi.org/10.1007/s10666-014-9404-8
31. Kuter S, Weber G-W, Akyürek Z, Özmen A (2015) Inversion of top of atmospheric reflectance values by conic multivariate adaptive regression splines. Inverse Probl Sci Eng 23(4):651–669. https://doi.org/10.1080/17415977.2014.933828
32. Kartal-Koç E, Iyigun C, Batmaz I, Weber G-W (2014) Efficient adaptive regression spline algorithms based on mapping approach with a case study on finance. J Glob Optim 60(1):103–120. https://doi.org/10.1007/s10898-014-0211-1

33. Çevik A, Weber G-W, Eyüboğlu BM, Oğuz KK (2017) Voxel-MARS: a method for early detection of Alzheimer's disease by classification of structural brain MRI. Ann Oper Res 258(1):31–57. https://doi.org/10.1007/s10479-017-2405-7
34. Özmen A, Yılmaz Y, Weber G-W (2018) Natural gas consumption forecast with MARS and CMARS models for residential users. Energy Econ 70:357–381. https://doi.org/10.1016/j.eneco.2018.01.022
35. Roy SS, Pratyush C, Barna C (2018) Predicting ozone layer concentration using multivariate adaptive regression splines, random forest and classification and regression tree. Adv Intell Syst Comput 634:140–152. https://doi.org/10.1007/978-3-319-62524-9_11
36. Nieto PJG, Álvarez JCA (2014) Nonlinear air quality modeling using multivariate adaptive regression splines in Gijón urban area (Northern Spain) at local scale. Appl Math Comput 235:50–65. https://doi.org/10.1016/j.amc.2014.02.096
37. Shahraiyni TH, Sodoudi S (2016) Statistical modeling approaches for PM10 prediction in urban areas: a review of 21st-century studies. Atmosphere 7(2):15. https://doi.org/10.3390/atmos7020015
38. Bai L, Wang J, Ma X, Lu H (2018) Air pollution forecasts: an overview. Int J Environ Res Public Health 15(780):1–44. https://doi.org/10.3390/ijerph15040780
39. Salford Systems Data Mining and Predictive Analytics Software Modeler, SPM Version 8.0 (2016). Salford Systems, San Diego, CA
40. SPSS IBM Statistics. https://www.ibm.com/analytics/data-science/predictive-analytics/spss-statistical-software. Accessed 15 July 2019
41. Wolfram Mathematica system. http://www.wolfram.com/mathematica/. Accessed 15 July 2019
42. Steinberg D, Golovnya M (2007) CART 6.0 user's guide. Salford Systems, San Diego
43. Death G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2
44. Wu X, Kumar V (2009) The top ten algorithms in data mining. Chapman & Hall/CRC, Boca Raton
45. Izenman J (2008) Modern multivariate statistical techniques: regression, classification, and manifold learning. Springer, New York
46. Burnham KP, Anderson DR (2002) Model selection and inference: a practical information-theoretic approach, 2nd edn. Springer, New York
47. Ljung GM, Box GEP (1978) On a measure of lack of fit in time series models. Biometrika 65:297–303. https://doi.org/10.1093/biomet/65.2.297
48. De Gooijer JG, Kumar K (1992) Some recent developments in non-linear time series modelling, testing, and forecasting. Int J Forecast 8:135–156. https://doi.org/10.1016/0169-2070(92)90115-P
49. Wilks DS (2011) Statistical methods in the atmospheric sciences, 3rd edn. Elsevier, Amsterdam
50. Dockery DW, Pope CA (1994) Acute respiratory effects of particulate air pollution. Annu Rev Public Health 15:107–132. https://doi.org/10.1146/annurev.pu.15.050194.000543
51. Yin P, He G, Fan M, Chiu KY, Fan M, Liu C, Xue A, Liu T, Pan Y, Mu Q, Zhou M (2017) Particulate air pollution and mortality in 38 of China's largest cities: time series analysis. Brit Med J 356:j667. https://doi.org/10.1136/bmj.j667
52. Katsouyanni K, Touloumi G, Spix C, Schwartz J, Balducci F, Medina S, Rossi G, Wojtyniak B, Sunyer J, Bacharova L (1997) Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: results from time series data from the APHEA project. Brit Med J 314:1658–1663. https://doi.org/10.1136/bmj.314.7095.1658
53. European Environment Agency (2017) Air quality in Europe—2017 report, EEA Report 13. https://www.eea.europa.eu/publications/air-quality-in-europe-2017. Accessed 15 July 2019
54. European Environment Agency (2018) Air quality in Europe—2018 report, EEA Report 12. https://www.eea.europa.eu//publications/air-quality-in-europe-2018. Accessed 15 July 2019
55. National System for Environmental Monitoring, Bulgaria (2013). http://eea.government.bg/en/nsmos/index.html. Accessed 15 July 2019
56. Executive Environment Agency (ExEA), Bulgaria. http://eea.government.bg/en Accessed 15 July 2019
57. Air Quality Guidelines for Europe (2000) 2nd edn, World Health Organization (WHO), Regional Office for Europe, Copenhagen. http://apps.who.int/iris/handle/10665/107335. Accessed 15 July 2019
58. Regional Inspectorate of Environment and Water—Ruse, Reports on the state of the environment (2011–2016). http://www.riosv-ruse.org/doklad-za-sastoyanieto-na-okolnata-sreda.html. Accessed 15 July 2019 **(in Bulgarian)**
59. RIOSV Pernik: Report on the state of air quality (2010–2014). http://pk.riosv-pernik.com/index.php?option=com_content&view=category&id=74:revisheniq&Itemid=28&layout=default (in Bulgarian). Accessed 15 July 2019
60. Ruse Historical Weather. https://www.worldweatheronline.com/ruse-weather-history/ruse/bg.aspx. Accessed 15 July 2019
61. Pernik Historical Weather. https://www.worldweatheronline.com/pernik-weather-history/pernik/bg.aspx. Accessed 15 July 2019
62. ALADIN Project for weather forecasts, Bulgaria (2019). http://www.weather.bg/0index.php?koiFail=cities1&lng=1&ci=Ruse&gr=Ruse. Accessed 15 July 2019