Assaying SARIMA and generalised regularised regression for particulate matter PM10 modelling and forecasting

Snezhana Gocheva-Ilieva* and Atanas Ivanov

Department of Applied Mathematics and Modelling,
Faculty of Mathematics and Informatics,
University of Plovdiv 'Paisii Hilendarski',
24 Tsar Asen str., 4000 Plovdiv, Bulgaria
Email: snegocheva@gmail.com
Email: aivanov_99@yahoo.com
*Corresponding author

Abstract: Two different predictive modelling approaches – classical SARIMA time series methodology and the new Generalised PathSeeker (GPS) regularised regression method, supported by stochastic gradient boosting trees, RuleLearner and other data mining techniques - are used to examine the concentration of particulate matter PM10 in the town of Kardzhali, Bulgaria. Empirical models are developed to simulate and forecast pollution levels based on hourly PM10 data from 1 January 2011 to 28 February 2014 in dependence on six meteorological variables. The constructed models have been used for 5-days-ahead hourly forecasts, compared to actual data from 1 to 5 March 2014. The obtained SARIMA and GPS models fit very well to historical data with coefficients of determination R^2 = 90% and 82% and root mean square error RMSE = 0.114 and 0.151, respectively. In forecasting, the GPS models outperform SARIMA approach. This could be explained by the preliminary classification provided by the data mining techniques and cross-validation procedure.

Keywords: air pollution; particulate matter PM10; seasonal ARIMA; generalised PathSeeker regularised regression; stochastic gradient boosting; data mining; forecasting; environmental pollution.

Reference to this paper should be made as follows: Gocheva-Ilieva, S. and Ivanov, A. (2019) 'Assaying SARIMA and generalised regularised regression for particulate matter PM10 modelling and forecasting', *Int. J. Environment and Pollution*, Vol. 66, Nos. 1/2/3, pp.41–62.

Biographical notes: Snezhana Gocheva-Ilieva is a graduate of Mathematics and obtained her MSc in Computational Mathematics from the Sofia University 'St. Kliment Ohridski', Sofia, Bulgaria in 1973. She completed her PhD in Physics and Mathematics from the 'Taras Shevchenko' National University of Kyiv, Ukraine, in 1981. In 2016 she defended a scientific degree of Doctor of Sciences in Mathematical modelling and application of mathematics at the University of Plovdiv 'Paisii Hilendarski', Bulgaria. She is currently working as a Full Professor of Applied Mathematics at the same university. Her research interests include various fields of mathematical modelling in physics and engineering, modelling in environmental science, applied computational statistics and predictive data mining techniques.

Atanas Ivanov is a graduate of Mathematics in 2010 and obtained his PhD in Mathematical Modelling and Application of Mathematics in 2015 at the University of Plovdiv 'Paisii Hilendarski', Bulgaria. His research is focused on time series modelling of air quality using the stochastic and modern intelligent data mining methods.

1 Introduction

Recently the European Environment Agency (EEA) published its report on air quality in European countries for the period 2015 (EEA, 2017). It indicated that the most polluted urban areas are located in Eastern Europe countries, including Bulgaria, and the levels of the main pollutants such as particulate matter PM10, PM2.5, ground level ozone and nitrate oxides systematically exceed the prescribed European limits. Their influence is crucial for public health and particularly that of small children and the elderly (Dockery and Pope, 1994; Davalos et al., 2017). This calls for special attention to be devoted to studies pertaining to these pollutants in affected regions, including continuous monitoring and processing of the data.

In Bulgaria, about 12 of the key air pollution indicators are systematically monitored by 36 automated stations run by the Executive Environment Agency in accordance with European legislation, directives and air quality standards (Directives on Ambient Air, 2008; Air Quality Standards, 2013). The accumulated observation data are processed daily and registered exceedances of these pollutants over the accepted limits are published in (ExEA, 2018). Most environmental statistics deal with modelling air quality and air pollution. Of the applied parametric methods, the most popular for times series analysis are the autoregressive integrated moving average (ARIMA) methods, also known as stochastic Box-Jenkins methodology (Box et al., 1994). In Lima et al. (2009), both total suspended particles and particles with a diameter equal or smaller than 10 μm (PM10) are analysed using the seasonal ARIMA approach. The Box-Jenkins methodology and multivariate analysis have been applied (Liu, 2009) for simulation and forecasting of the daily average PM10 concentrations. In Al-Madfai et al. (2010) three models have been built using univariate Box-Jenkins, ARIMA transfer function and variability decomposition methods, which are compared and used for analysing the multi-year maximum daily PM10 concentration. A combination of SARIMA and support vector machine was applied in (Lee et al., 2017) for PM10 and ozone modelling.

Many new statistical techniques and algorithms for modelling air pollution data are being developed and applied as an alternative to parametric methods. Data mining techniques are intelligent methods for data-based modelling using specially developed high-performance computational algorithms, including for large data sets ('big data'). Among the methods and techniques of data mining are: neural networks, decision trees, random forests, support vector machines, etc. (Nisbet et al., 2009). These techniques are often combined and compared with classical methods. Ul-Saufie et al. (2013) examine the concentrations of PM10 in order to achieve short-time forecasting by combining multiple regressions model and feed forward back propagation neural network models with principle component analysis. Biancofiore et al. (2017) have conducted a study to forecast the average daily PM10 and PM2.5 concentrations of one to three days in the urban area of the Adriatic coast using data for three years. The analysis is made using multiple linear regressions, neural networks with and without recursive architecture. Ganesh et al. (2017) predicted the AQI air quality index in Houston and Los Angeles using data on concentrations of NO_2, CO, O_3, PM2.5, SO_2 and PM10 in the period from 2010 to 2016 using artificial neural networks, multiple linear regression and vector regression. Wang et al. (2015) used hybrid artificial neural network and a hybrid support vector machine by revising the error term of the traditional methods to forecast the daily PM10 and SO_2 concentrations. More data mining techniques and model approaches are also used for air pollutants, including random forest, decision tree, support vector machine, ANN and more (Grange et al. 2017; Noori et al., 2010; Moazami et al., 2016; Stoimenova et al. 2017; Whalley and Zandi, 2016). Recent results for modelling PM10 can be found in Dotse et al. (2018), Shahraiyni and Sodoudi (2016), Siwek et al. (2011) and Zheleva et al. (2017), including an application of GPS regularised regression in Ivanov and Gocheva-Ilieva (2013) and the literature cited therein.

The goal of this paper is to analyse PM10 data for modelling and forecasting of time series in relation to changes in meteorological variables. Two approaches are used and compared – classical SARIMA and GPS regularised regression, in combination with powerful data mining techniques such as stochastic gradient boosting trees (known as TreeNet), importance sampled learning ensembles (ISLE) and RuleLearner (Friedman 2001, 2012; Friedman and Popescu, 2003, 2005). As a case study the air pollution of the town of Kardzhali, Bulgaria is modelled on the basis of hourly data for particulate matter PM10 and six meteorological variables over a period of three years and two months. The novelty of the investigation is the application of GPS, taking into account new modelling aspects such as using the time-variables, including such of continuous type, data transformation, relatively long forecasted period, etc. Other factors affecting PM10 are not included in the models due to the lack of necessary data.

The main stages of the investigation are:

1 Description of study area, data and methods used.

2 Conducting time series analysis by the stochastic SARIMA models, accounting for periodicity (in hours) and dependence of PM10 emissions on six meteorological variables.

3 Conducting GPS models, enhanced by the data mining techniques for the considered data.

4 Analysis and estimation of the obtained models.

5 Application of models for 5-days (120 hours) ahead forecasting.

6 Comparison of the models and their performance characteristics based on the obtained results.

The models were built with the help of Salford Predictive Modeler software suite including (GPS), TreeNet, ISLE and RuleLearner (SPM, 2017), IBM SPSS 24 software package and Wolfram Mathematica.

2 Study area and data description

Data are collected for air pollution in the area of Kardzhali, located in the Eastern Rhodopes Mountain in South Bulgaria, 260 km from the capital Sofia. The town is located on both banks of the river Arda between two large artificial lakes. The exact coordinates of the town are: 41°39’N, 25°22’E, altitude of 275 m. Kardzhali has a humid moderate climate. Drought is characteristic for the town throughout the year with some rainfall during winter. Summers are hot and winters are cold. Its population is around 45,000 people. There is little road traffic as it is located away from busy highways. There are two main sources of pollution in Kardzhali – households (use of solid fuels for heating) and several large heavy industry factories (heavy metal processing plants and others), which release harmful emissions into the atmosphere. There is no immediate pollution from other nearby towns or cities.

The analysed data for the concentrations of PM10 in the town of Kardzhali have been collected over a period of three years and two months – from 1 January 2011 to 28 February 2014 taking into account hourly data. Data are expressed in units of mass concentration in micrograms per cubic metre. The influence of the following six meteorological variables is also taken into account: temperature of ground air (TEMP), atmospheric pressure (PRESS), relative humidity (UMR), wind speed (WS), wind direction in degrees (WD) and sun radiation (RADST). The data used in the modelling process were taken from n = 27,720 (hours) cases. Table 1 shows the main descriptive statistics for the seven observed time series, assuming the PM10 pollutant is a dependent variable and the six meteorological variables are independent. Missing measured data are 1.6% for PM10 and less than 0.63% for meteorological variables. All missing values were replaced by using linear interpolation.

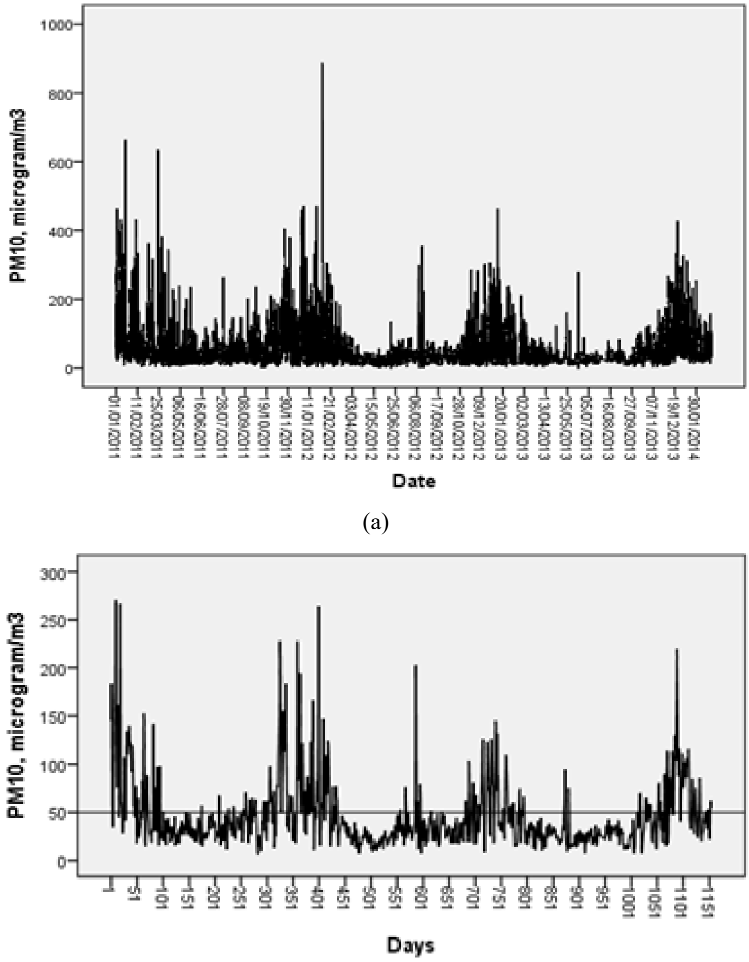
Table 1 Descriptive statistics for the observed time series PM10 and the six meteorological variables for city of Kardzhali, Bulgaria[1]

Variable	*Minimum*	*Maximum*	*Mean*	*Median*	*Std. dev.*	*Skewness*[2]	*Kurtosis*[2]
PM10 (μg/m^3)	1	887.0	45.25	30.50	47.441	3.827	24.407
TEMP (°C)	–15.3	38.9	12.58	12.38	9.922	0.146	–0.776
PRESS (mbar)	948.6	1000.6	978.97	978.75	6.756	–0.078	0.375
WS (m/s)	0.1	16.0	1.05	0.74	0.843	1.740	7.169
UMR (%)	4.13	96.8	63.39	58.03	30.194	–0.131	–1.578
RADST (W/m^2)	0.5	996.5	165.05	10.95	255.898	1.598	1.344
WD (degree)	22.5	315.0	169.80	157.50	47.738	0.526	–0.228
trPM10	0.36	3.64	2.5045	2.4901	0.360	–0.169	1.119

Notes: [1]n = 27720. [2]Std. err. skewness = 0.015, std. err. Kurtosis = 0.029.

Figure 1(a) shows a sequence plot of the raw data in hours. From Table 1 and this figure it is observed that the time series of PM10 has high values with an average of 45.25 μg/m^3. Maximum concentration is generally observed in the winter months from October to April reaching up to 887 μg/m^3. Figure 1(b) illustrates the aggregated daily average values for the same data. Systematic exceedances of the daily average PM10 concentrations over the European and national daily pollution limit of up to 4-5 times are observed. The total number of daily exceedances for the considered time period is 307, or 26.8%. This could be classified as a serious ecological problem for the town. The pollution may be attributed to the extensive use of solid fuels by households and in industrial manufacturing, which is the main risk and a very harmful factor to human health in this urban area.

Figure 1 (a) Observed data for PM10, measured by hours in the town of Kardzhali
(b) The PM10 daily mean concentrations for the same period



(a)

(b)

Note: The horizontal line represents the official PM10 daily limit upper value of 50 μg/m^3.

3 Methods

3.1 SARIMA methodology

The well-known general classes of stochastic models ARIMA and its seasonal extension SARIMA, introduced by Box and Jenkins in 1970 comprise simultaneously a number of parameters for common and more accurate representation of time series in a stochastic way (see Box et al., 1994). The method also allows the consideration of the dependences of times series on other input series, taking into account, for example meteorological, climatic and other factors.

In the univariate case, when time series Y depends only on time t, the general form of the multiplicative SARIMA(p, d, q)(P, D, Q)$_{24}$ model is represented by nonnegative integer parameters p, d, q, P, D, Q, s, where p, P are autoregressive non-seasonal and seasonal processes (AR and SAR); d, D – non-seasonal and seasonal trend processes (I), identified by a differencing procedure until the time series becomes stationary; q, Q – moving average non-seasonal and seasonal processes (MA); s – the number of seasons or period of periodical patterns in time series behaviour.

Let the time series values Y_t are known at time t = 1, 2, 3, ..., n, where n is the total number of observations and B denotes the backward lag distance operator, defined as $BY_t = Y_{t-1}$, …, $B^kY_k = Y_{n-k}$. The univariate multiplicative SARIMA(p, d, q)(P, D, Q)$_s$ model has the form (Box et al., 1994):

$$\phi_p(B)\Phi_P\left(B^s\right)(1-B)^d\left(1-B^s\right)^D Y_t = \theta_q\left(B\right)\Theta_Q\left(B^s\right)a_t + c,\ a_t \sim WN\left(0, \sigma^2\right) \qquad (1)$$

where ϕ_p, Φ_P are difference polynomials for non-seasonal and seasonal autoregressive parameters, θ_q, Θ_Q are non-seasonal and seasonal moving average parameters, $a_t \sim WN(0, \sigma^2)$ is a white noise term, c is a constant. Usually, it is assumed that the time series is stationary (d = 0). If non-stationary, it may be transformed into stationary by some degree of initial differencing, denoted by the differencing operator (1 – B) in (1).

When time series Y is influenced by multiple input time series X_1, X_2, …, X_r, defined for the same time values, a multiple-input transfer function model can be constructed (Box et al., 1993):

$$Y_t = \sum_{j=1}^{r} \frac{\omega_j(B)}{\delta_j(B)} B^{b_j} X_{jt} + \frac{\theta_a(B)}{\Phi_a(B)} a_t \qquad (2)$$

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

The assumptions for applying the ARIMA type methods include the condition of normality of the time series, small residuals as a white noise, using statistical tests (Ljung-Box or others) to check that the model residuals have no autocorrelation, etc. If the assumption of normality is violated, for improving the distribution and stabilising the variance of the data for PM10, the following Yeo-Johnson power transformation could be applied:

$$trY(\lambda, Y) = \begin{cases} \left\{(Y+1)^{\lambda} - 1\right\} / \lambda & Y \geq 0, \lambda \neq 0 \\ \log(Y+1) & Y \geq 0, \lambda = 0 \\ -\left\{(-Y+1)^{2-\lambda} - 1\right\} / (2-\lambda) & Y < 0, \lambda \neq 2 \\ -\log(-Y+1) & Y < 0, \lambda = 2 \end{cases}, \quad \lambda \in [-2, 2] \qquad (3)$$

where Y is the initial variable (here PM10) and trY is the transformed variable, λ is a parameter (Yeo and Johnson, 2000).

3.2 Introduction to GPS regularised regression and used data mining techniques

The GPS is a new statistical method developed by Friedman (2012). Among its main advantages is the capability to process efficiently large matrices of continuous or binary data, to work well with highly correlated predictors and at a presence of high variance.

Let us assume that m independent variables (predictors) ($X_1, \dots, X_m$) are used to fit the dependent variable Y. The GPS model has the form of ordinary linear regression:

$$\hat{Y} = a_0 + a_1X_1 + a_2X_2 + \ldots + a_mX_m,\ \varepsilon_t = Y_t - \hat{Y}_t,\ t = 1, \ldots, n \tag{4}$$

where $\mathbf{a} = (a_0, a_1, \dots, a_m)$ is the vector of the regression coefficients (parameters) and ε_t are the residuals (or error terms) of the model. However, it must be noted that practically, instead of the initial predictors, a large number of secondary variables are constructed using additional data mining techniques, as described below.

The standard regularised modelling aims to find a regression equation (4) for fitting to data by solving the following optimisation problem for the regression coefficients:

$$\hat{\mathbf{a}}(\eta) = \underset{\mathbf{a}}{\arg\min}\left[L(\mathbf{a}) + \eta P(\mathbf{a})\right] \tag{5}$$

where $L(\mathbf{a})$ is the empirical loss function, selected among different error criteria (usually a mean squared error (MSE) with $L = \sum_{t=1}^{n}[Y_t - \hat{Y}_t]^2/n)$, $P(\mathbf{a})$ is a penalty function and $\eta > 0$ is the regularisation parameter. In the extended power family the penalty function is generalised to the form $P_\gamma(\mathbf{a}) = \sum_{j=0}^{m}\left|a_j\right|^{\gamma}, 0 \le \gamma \le 2$ (Friedman, 2012). For $\gamma = 1$ one obtain the Lasso regression, which allows to introduce variables sparingly and generate reasonably sparse solutions (Tibshirani, 1996). For $\gamma = 2$ it is a ridge regression (Hoerl and Kennard, 1970), which contributes to estimates stabilisation in the presence of extreme multicollinearity. The difference between Lasso and ridge regression is that in ridge regression, as the penalty is increased, all parameters are reduced while still remaining non-zero, while in Lasso, increasing the penalty will cause more and more of the parameters to be driven to zero. Further extension in terms of penalty function is represented by the elastic net family (Zou and Hastie, 2005; Friedman, 2012):

$$P_\beta(\mathbf{a}) = \sum_{j=1}^{m}(\beta - 1)a_j^2/2 + (2 - \beta)\left|a_j\right|, \quad 1 \le \beta \le 2 \tag{6}$$

where β is the coefficient of elasticity, extended further in $0 \le \beta < 1$.

With the GPS method, the problem (5)–(6) is generalised and the calculation complexity has been resolved through sequential path seeking directly in the parameter space under a given penalty $P(\mathbf{a})$ without solving the optimisation problem at each step. The meta-logic of the GPS algorithm and examples are given in (Friedman, 2012).

The GPS regularised regression is realised as a very fast forward stepping algorithm with a flexible variable selection procedure. At every step a new variable, selected to satisfy a given set of criteria is added, or the coefficient of some model variable is adjusted. Also, as in other machine learning techniques, the models are data driven and cross-validation procedures are preferred for model selection. To this end, the sample at hand is randomly divided in two non-overlapping subsamples – a learning sample and a test sample. The model is built on the learning sample and subsequently its fitting ability is estimated by the test sample.

The quality of models is evaluated using a commonly used goodness-of-fit measures for learning and test samples. Within regression the following are calculated for both learning and test samples: the coefficient of determination R^2, root mean squared error (RMSE), mean squared error (MSE), mean absolute deviation (MAD) and other.

Despite its advantages, the GPS method has some limitations. As a linear regression technique it is sensitive to data distribution. The method does not provide automatic discovery of nonlinearities, interactions between predictors, or a missing values handling feature. In practice, the usage of GPS is highly efficient in combination with the powerful predictive data mining techniques, such as stochastic gradient boosting, ISLE and RuleLearner. The basic one of them is the stochastic gradient boosting, which general computational approach is also known as TreeNet (Salford Predictive Modeler Users Guide, 2013). This method is applied for pre-processing the data in order to obtain more adequate predictors, based on classification procedure by generating a family of trees. TreeNet is designed to handle both regression and classification problems. The original model, produced by TreeNet algorithm is a sequence of hundreds or even thousands of small trees (normally with two to six terminal nodes). At any current stage of construction of the model, it is drawn a random subsample from the learning data and the successive tree is built for the prediction residuals of the preceding tree. In essence every tree can be considered as a new variable transformation represented by a continuous variable as a function of inputs. This way, performing consecutive boosting computations on independently drawn subsamples of observations one obtain a stochastic gradient boosting (i.e., TreeNet) model. A simple example of obtaining a binary classification tree is as follows. Let a subsample of observations of the dependent variable Y and X_1 be its predictor, for example X_1 is air temperature. All observations measured at a temperature of less than 5°C are classified into one node and the others in another. We get a tree with two terminal nodes. The predicted model values are the arithmetic means of Y-values in each terminal node. Splitting procedure can continue with the same or another predictor and its values for obtaining more nodes. Taking the residuals as a new dependent variable one can produce a new tree, to reduce the overall model error, etc. The sum of these trees gives an example of a TreeNet gradient boosting trees model for the selected training subsample. The more details on the TreeNet algorithm are given in (Friedman, 2001).

In fact, many of the obtained trees are usually equal or have very similar structure. The compression of the TreeNet model can be performed using the ISLE algorithm by removing redundant trees. The coefficients of the linear model (4) are then adjusted by the GPS. In addition, the RuleLearner algorithm could be applied as a post-processing technique which selects the most influential subset of nodes, thus further reducing model complexity (Friedman and Popescu, 2005). Different combinations of these techniques can be carried out for model selection to obtain the best GPS model.

3.3 Model selection and evaluation criteria

When evaluating the quality of constructed SARIMA and GPS models, we will use the classical measures for the goodness of fit such as RMSE, R^2 and mean absolute percentage error (MAPE):

$$R^2 = \frac{\sum_{t=1}^{n}\left(\hat{Y}_t - \bar{Y}\right)^2}{\sum_{t=1}^{n}\left(Y_t - \bar{Y}\right)^2}, RMSE = \sqrt{\sum_{t=1}^{n} \varepsilon_t^2 / n}, MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|, \varepsilon_t = Y_t - \hat{Y}_t \qquad (7)$$

where $\hat{Y}_t$ is the predicted value of Y_t at time moment t, $\bar{Y}$ is a mean value, ε_t is the residual. Among the preferred models is the one with a maximum R^2 and minimum error estimates RMSE and MAPE. For GPS models, the usual 10-fold cross-validation procedure is applied. When some models have almost identical characteristics, the parsimonious principle is used (Box et al., 1994).

It is also important to validate and compare the models with respect to known holdout data. In this paper, we validate the models by calculating 5-days-ahead forecasts using observations from the same series, which were not used in the model.

4 Results and discussion

We will describe the basic results obtained using the best models built by means of SARIMA and GPS enhanced by the upper mentioned data mining techniques.

4.1 Building and analysing SARIMA models of PM10 concentrations

The SARIMA models are constructed according to the following scheme:

1 Check the time series for trends, weak stationarity and invertibility.

2 Perform the transformation of the dependent variable PM10 to improve distribution and stabilise the variance. Denote the result by trPM10. Transform the variable WD for wind direction into continuous.

3 Plot and test autocorrelation function (ACF) and partial ACF (PACF) of PM10 and trPM10 to determine seasonality s and identify the limits of ARIMA model parameters.

4 Subtract the average value from the transformed trPM10 series; denote the result with m_trPM10.

5 Build and explore ARIMA/SARIMA models for m_trPM10:

   a only dependent on time

   b with all predictors

   c with the most important predictors.

6 Evaluate and analyse the adequacy of the models regarding performance measures (7), error residuals, assessing whether the residuals exhibit autocorrelation using Ljung-Box test, etc. Choose the best models.

7 Apply the selected models to forecast PM10 concentrations for 5 days ahead (120 hours) period.

Following this general scheme we calculated the unit root ADF test for PM10 time series. The resulting significance is $1.10^{-7} < 0.05$, hence the series PM10 does not have a linear trend. From the meteorological variables only for the PRESS variable, a linear trend was found with coefficient equal to 0.6809. The rest of variables have no trend. Verifications for weak stationarity and invertibility of the processes showed. 'True' for all variables in the analysis. The upper results have been obtained using the respective built-in functions of Wolfram Mathematica software, version 10.2.

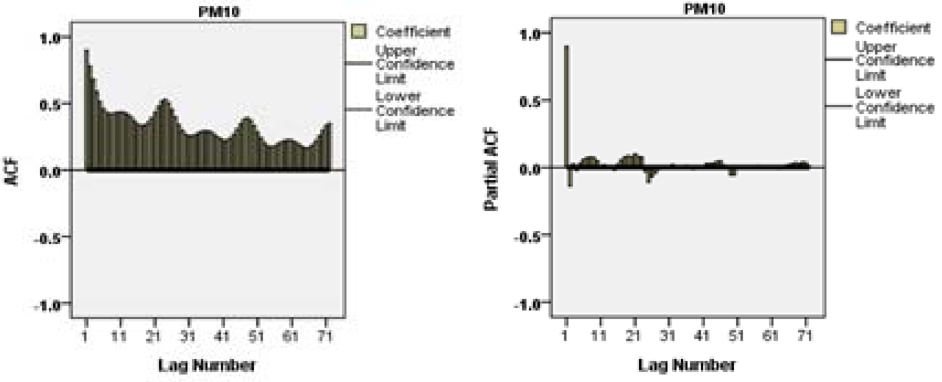
A basic assumption for applying Box-Jenkins ARIMA methodology is the normal or near-normal distribution of the dependent variable and relevant predictors (Wilks, 2011). Using coefficients of skewness and kurtosis of PM10 from Table 1 and divided them by their corresponding standard errors we obtain absolute values higher than 2.5 for PM10 and for a part of the meteorological variables. This indicates that the distribution of the variables is not normal. The same result was obtained by the Kolmogorov-Smirnov and other tests of normality. For improving the distribution of PM10, the optimal value of λ in (3) was chosen to be $\lambda = -0.2$ by using a simple procedure of attempts and checking the Kolmogorov-Smirnov test of normality. Corresponding descriptive statistical indices for the transformed variable *trPM*10 are given in the last row of Table 1.

In order to account for the periodicity of the WD this variable was transformed into a new variable WDI by the expression

$$WDI = 1 + \sin(WD + \pi / (k - 1)) \tag{8}$$

where $k = 16$ is the number of the main wind directions from 0 to 360 degrees.

Figure 2 (a) ACF of PM10 (b) PACF of PM10 (see online version for colours)



(a) (b)

The next step is to draw ACF and PACF of the PM10 to identify seasonality and the limits of ARIMA modelling parameters. Figures 2(a) and 2(b) shows the ACF and PACF of the PM10 series. In ACF there is a clear cyclicality of 24 lags, which suggests to set seasonality s = 24. In PACF of PM10 this seasonality is confirmed and the lack of a jump equal to 1 at lag 1 confirms that the series have no linear trend, which we checked with the ADF test. The behaviour of PACF assumes the following intervals for the best tentative SARIMA models: p from 1 to 3 and q from 6 to 12. Seasonal parameters *P* and

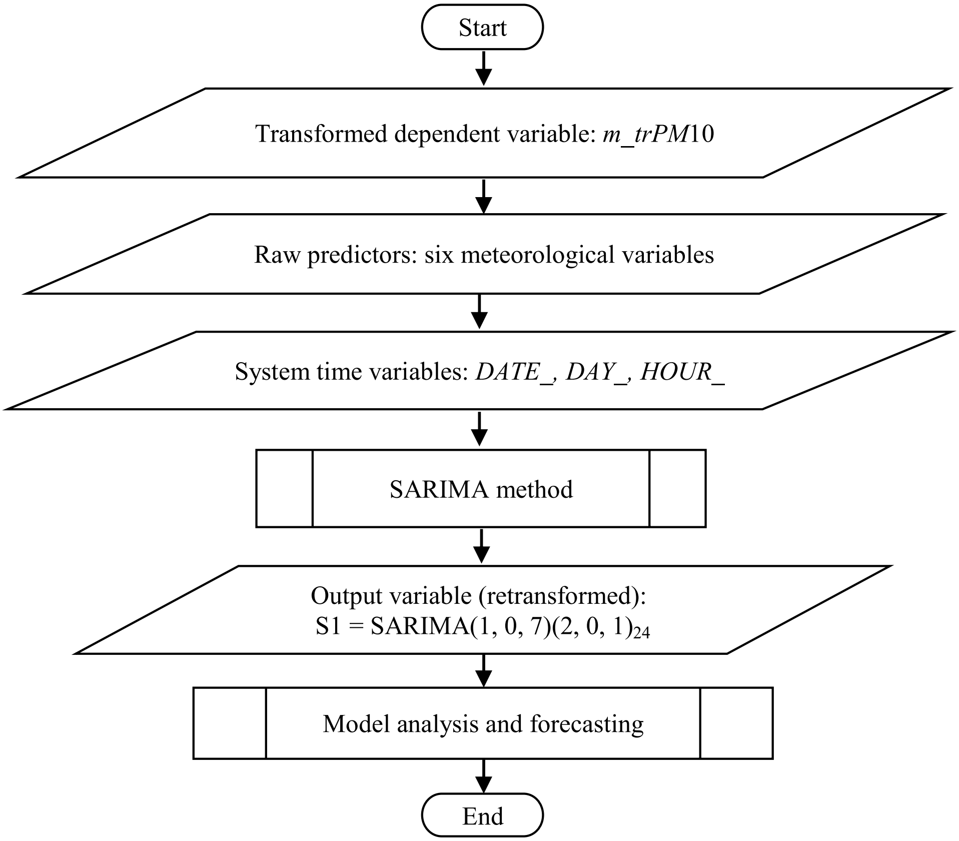
Q are difficult to assess on these plots, but the lack of large fluctuations indicates they do not have large values (Tabachnik and Fidell, 2007).

Following the general procedure we calculate a new dependent variable:

$$m\_trPM10 = trPM10 - 2.5045 \tag{9}$$

where 2.5045 is the average value of *trPM*10 (see Table 1). This suggests that the SARIMA models could be built without a constant. Generating time series models with the ARIMA algorithm requires the inclusion of system time variables, such as *DATE_*, *DAY_* and *HOUR_* in the case of hourly data, when using SPSS.

Figure 3 Block diagram for SARIMA model S1


Start
Transformed dependent variable: *m_trPM*10
Raw predictors: six meteorological variables
System time variables: *DATE_*, *DAY_*, *HOUR_*
SARIMA method
Output variable (retransformed):
S1 = SARIMA(1, 0, 7)(2, 0, 1)$_{24}$
Model analysis and forecasting
End

By varying the parameters (*p*, 0, *q*), (*P*, 0, *Q*) we obtained the best indices for SARIMA(1, 0, 7)(2, 0, 1)$_{24}$ models. The autoregressive component ($p = 1$) indicates the strongest influence of the previous one hour on every current level of PM10 emission.

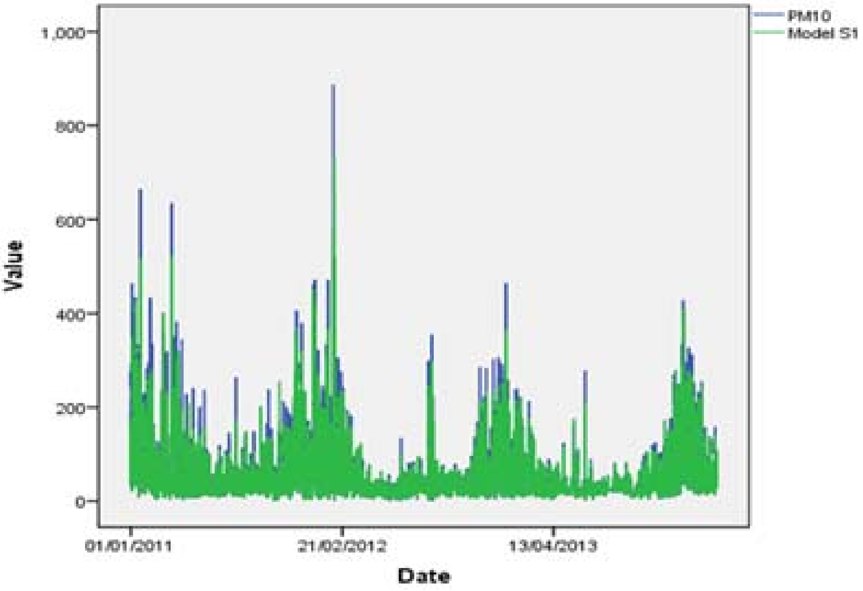
Three SARIMA models were selected with different predictors. The obtained performance statistics are given in Table 2. The values of coefficient of determination R^2 indicate very good fit of all the models to the observed data. It was established that in SARIMA model the inputs series *PRESS*, *TEMP* and *WS* are more important and other meteorological factors did not significantly affect the values of PM10. With almost identical statistical indices, model S1 was chosen as the best, as it is the simplest with only three out of six meteorological predictors.

Table 2 Summary statistics of the obtained best SARIMA(1, 0, 7)(2, 0, 1)$_{24}$ models for the PM10 concentrations in the town of Kardzhali, Bulgaria[1]

Model	*Model fit statistics*			*R^2 of 5 days ahead (120 hours) forecast*
	R^2	*RMSE*	*MAPE*	
S1	0.900	0.114	148.397	0.507
S2	0.900	0.114	148.768	0.512
S3	0.888	0.120	147.524	0.195

Notes: [1]Predictors for model S1: *PRESS*(2, 1, 0), delay 1; *TEMP*(2, 0, 1); *WS*(0, 0, 1). Predictors for model S2: *PRESS*(2, 1, 0), delay 1; *TEMP*(2, 0, 1); *WS*(0, 0, 1); *RADST*(0, 0, 0), delay 1; *UMR*(0, 0, 0), delay 1; *WDI*(0, 0, 0), delay 1. Predictors for model S3: no predictors. All seasonal parameters are set to (0, 0, 0)$_{24}$.

Figure 4 Observed data of PM10 compared to fitted values by the best SARIMA model S1 (see online version for colours)



For the sake of clarity, Figure 3 shows a block diagram for the SARIMA model S1 with the used variables and the construction stages (see also Table 2).

Actual and fitted data by model S1 are compared in Figure 4. There is very good fit between the values of PM10 predicted by the model S1 and measured ones.

For model adequacy, it is important to study how well the model fits the initial time series. To this end usually one examines the residuals, calculated as a difference between predicted and observed responses. The calculated residual ACF and PACF for the selected SARIMA(1, 0, 7)(2, 0, 1)$_{24}$ model S1 on the first 24 lags are shown in Figure 5 with a 5% confidence interval. It is shown that they are not significantly different from zero. It may be inferred that the residuals are small enough and the model S1 is adequate.

Figure 5 Plots of ACF and PACF of the residuals of the SARIMA(1, 0, 7)(2, 0, 1)$_{24}$ model S1 with the three significant meteorological predictors

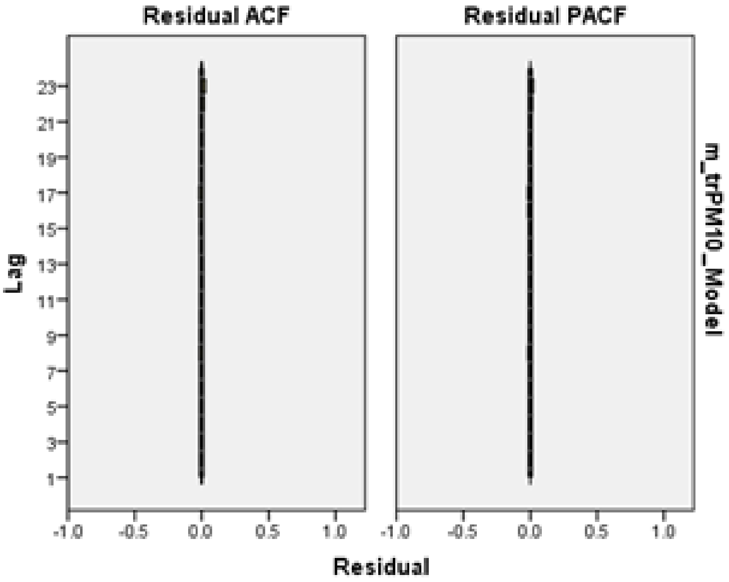
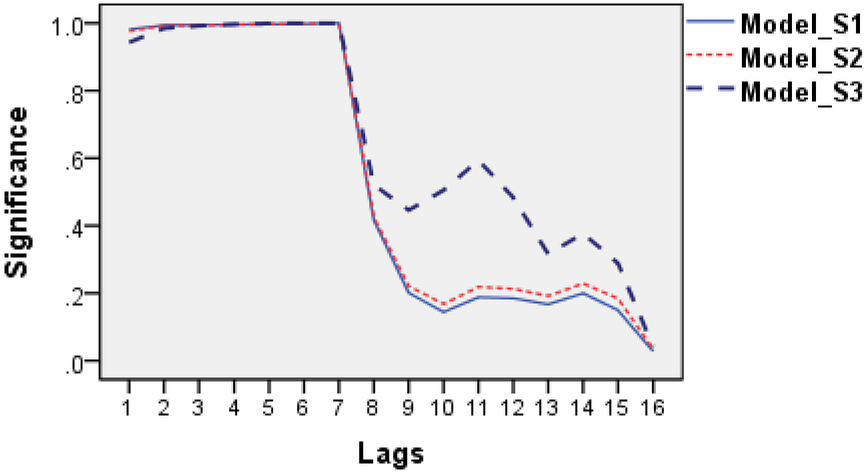


Figure 6 Plots of Ljung-Box ACF residual coefficients for SARIMA models S1, S2 and S3 (see online version for colours)



In order to assess whether the residuals show autocorrelation, we performed the Ljung-Box test for corresponding residual ACF. The results for the three models are shown in Figure 6. All coefficients are nonsignificant thus the series of residuals for the models S1, S2 and S3 exhibit no autocorrelation.

4.2 Construction and analysis of GPS regularised regression models of PM10

The construction of GPS models was carried out for *m_tr*PM10 as dependent variable. The six meteorological input series and time-variables *DAY*, *HOURS* and *DAYS_HOURS* were used as raw predictors. The variable *DAYS_HOURS* was defined as a continuous by the expression.

$$DAYS\_HOURS = round(DAY + (HOUR - 0.5) / 24, 3) \tag{10}$$

First, the initial (original) GPS regression model and the generation of the set of predictors by TreeNet are derived, by means of a sequence of small classification and regression trees. Then the GPS models, combined with additional techniques such as ISLE and RuleLearner and their combinations, are built. The validation of all models was performed using a standard 10-fold cross-validation procedure (Salford Predictive Modeler Users Guide, 2013). As a control settings there were used: maximum trees for TreeNet – 400 and 500, TreeNet and GPS loss function – least squares.

Seven obtained best models of dependent variable *m_tr*PM10 with their main statistics are given in Table 3. Models R1 to R5 were obtained with the raw predictors: *WS*, *TEMP*, *PRESS*, *UMR* and the three time variables – *DAYS_HOURS*, *HOUR* and *DAY*. Model R6 was calculated without the series *RADST* and *WDI*. Models R1 to R5 were built with maximum 500 trees in TreeNet and the last two models R6 and R7 – with maximum 400 trees.

Of all models in Table 3, the best indices have been found for model R5 of type GPS/ ISLE_RuleLearner_RawPredictors (IS_RL_RP). For this model, the coefficient of determination of the learning and test samples are higher, respectively 82.3%% and 81.1%. The elasticity of the model is Ridged Lasso (1.1).

Figure 7 shows a block diagram for the selected GPS model R5 with the variables used (see also Table 3).

The model selection was also based on the basis of the model accuracy. Relative predictive statistics of the best IS_RL_RP models with different raw predictors and maximum number of trees are presented in Table 4 with the R^2 of 5-days (120 hours) ahead forecasts. It is seen that the best performance is this of model R5.

The variable importance of the raw predictors in the best GPS models R5, R6 and R7 is given in Table 5. The order by strength of influence of the predictor variables is the same in all the three models and varies slightly. It is observed that the wind speed (*WS*) has the predominant role on the levels of the PM10 concentration. This could be explained by the conditions of the pollution dispersion at relatively weak wind speed with mean value 1.06 m/s and standard deviation 0.843 (see Table 1). The influence of the other meteorological series is mainly due to the air temperature (*TEMP*) and atmospheric pressure (*PRESS*). The GPS model includes substantially the two used time variables, reflecting the dependency on the time. These results correspond well to the results from the best SARIMA model S1. The sun radiation (*RADST*) and wind direction (*WDI*) have no significant influence on PM10.

Table 3 Pipeline GPS models and their basic statistics for the transformed series *m_trPM*10, for random learn and test samples, respectively[1]

Model	*GPS/RuleLearner model*	*Learn R^2*	*Learn N coef.*	*% Compress*	*Test R^2*	*Test N coef.*	*% Compress*	*Elasticity*
R1	Original (TreeNet)	0.735	500	0.0	0.707	500	0.0	
R2	RuleLearner	0.354	230	95.4	0.353	233	95.3	Ridge (2.0)
R3	RuleLearner_RawPredictors	0.818	688	86.2	0.801	625	87.5	Lasso (1.0)
R4	ISLE_ RuleLearner	0.354	230	95.4	0.351	233	95.3	Ridge (2.0)
R5	*ISLE_RuleLearner_RawPredictors*	*0.823*	*618*	*87.6*	*0.812*	*618*	*87.6*	*Ridged Lasso (1.1)*
R6	ISLE_RuleLearner_RawPredictors	0.809	560	86.0	0.797	560	86.0	Ridged Lasso (1.1)
R7	ISLE_RuleLearner_RawPredictors	0.808	578	85.6	0.795	532	86.7	Lasso (1.0)

Notes: [1]Predictors for models R1 to R6: *WS*, *TEMP*, *PRESS*, *UMR* and *DAYS_HOURS*, *HOUR*, and *DAY*. Predictors for model R7: all meteorological variables, *DAYS_HOURS*, *HOUR*, *DAY*.

Figure 7 Block diagram for the structure of the GPS model R5

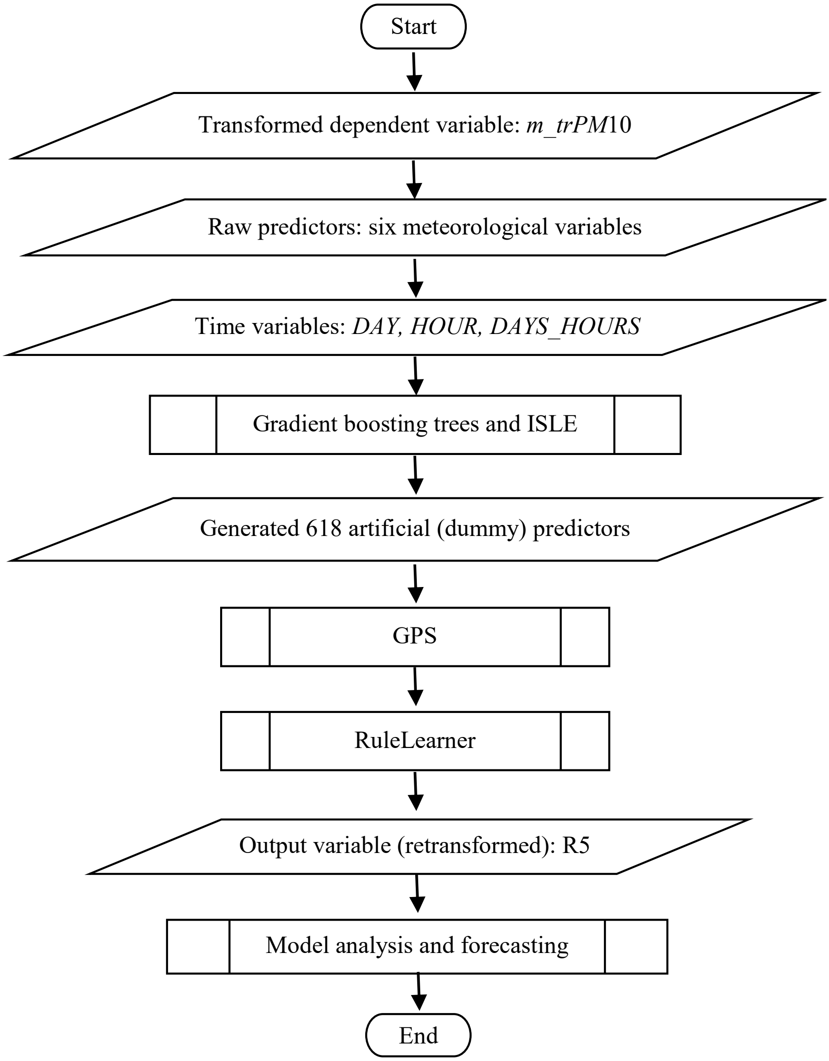

Start
Transformed dependent variable: *m_trPM*10
Raw predictors: six meteorological variables
Time variables: *DAY, HOUR, DAYS_HOURS*
Gradient boosting trees and ISLE
Generated 618 artificial (dummy) predictors
GPS
RuleLearner
Output variable (retransformed): R5
Model analysis and forecasting
End

Table 4 Summary statistics of the obtained best GPS models and their forecasting performance

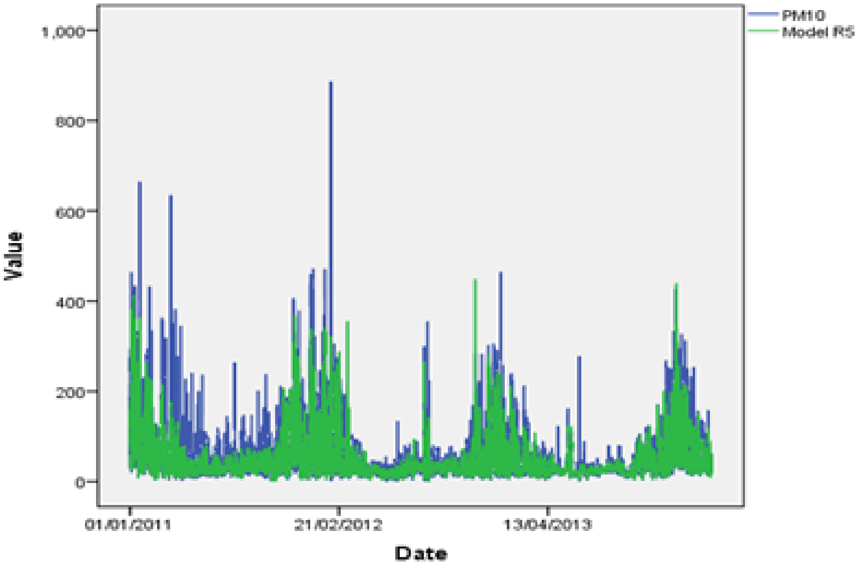
GPS model	*Model fit statistics*			*R^2 of 5 days ahead (120 hours) forecast*
	R^2	*RMSE*	*MAPE*	
R5	0.823	0.151	226.32	0.554
R6	0.806	0.158	220.24	0.556
R7	0.803	0.160	226.75	0.444

Figure 8 shows a comparison between the measured PM10 values and those predicted by the selected best GPS model R5.

Table 5 Relative variable importance of raw predictors in the selected three best GPS IS_RL_RP models

Variable	*Model R5*	*Model R6*	*Model R7*
DAYS_HOURS	100.0	100.0	100.0
WS, wind speed	77.33	79.01	81.54
TEMP, air temperature	62.83	64.70	66.37
HOUR	43.9	44.99	44.29
DAY	42.9	44.82	43.05
PRESS, pressure	41.0	42.33	43.13
UMR, relative humidity	31.7	31.00	30.75
RADST, sun radiation	–	–	20.01
WDI, wind direction	–	–	13.70

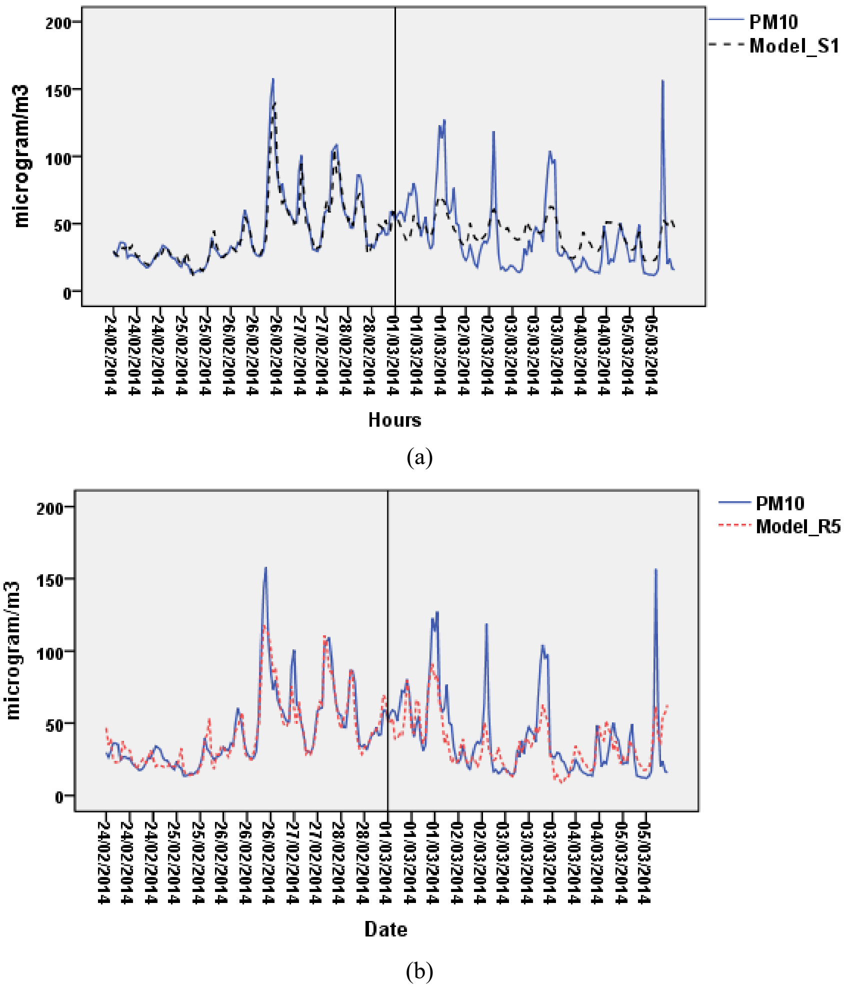
Figure 8 Observed hourly data of PM10 compared with fitted values by the best GPS/IS_RL_RP model R5 (see online version for colours)



4.3 *Forecasting results and comparison of the best models*

In this section we present some results demonstrating the application of the obtained SARIMA and GPS/RuleLearner models for a 5-days ahead hourly forecasting, based on data not used in the modelling process. To this end, the meteorological and time values for future 120 hours which are necessary for forecasting have been added to the predictor series and the predicted values of PM10 concentrations have been calculated, using the models. However, in real conditions future values of meteorological variables can be derived from 3 to 10 daily and hourly weather forecasts. For Bulgaria they are based on satellite and sensor data, processed by the ALADIN system and are updated daily every 12 hours [for Kardzhali see ALADIN (2018)].

Figure 9 Observed hourly data of PM10 compared with fitted values, (a) by the best SARIMA(1, 0, 7)(2, 0, 1)$_{24}$ model S1 (b) by the best GPS IS_RL_RP model R5 (see online version for colours)



(a)

(b)

Notes: On the left side of the vertical line – for 24 February to 28 February 2014 and on the right side – the forecasts for a 5-days-ahead period (from 1 March to 5 March 2014).

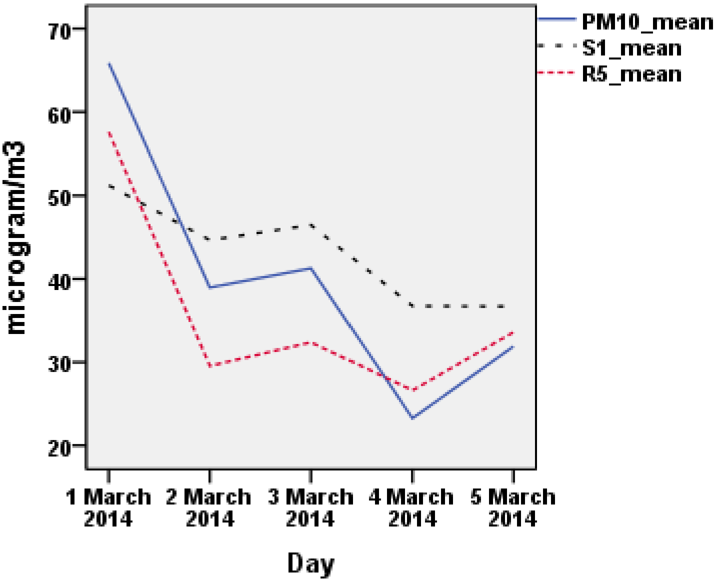
The retransformed forecasts by the best SARIMA models and GPS IS_RL_RP models, compared to the measured PM10 concentrations are presented in Figure 9.

For the forecasted 120 hours the goodness-of-fit R^2 statistics of the best SARIMA and GPS models are given in the last column of Table 2 and Table 4, respectively. The best forecasts are achieved by all GPS models with about R^2 = 55%.

Figure 9(a) shows that the SARIMA model S1 predicts to a very high degree the PM10 concentrations within the known historical type data, but for the out-of-sample data demonstrates moderate behaviour and the forecasts do not very adequately reflect the changes of the studied variable. Figure 9(b) for the best GPS model R5 predicts well the known data and shows better results in forecasting the 5-days out-of-sample data. The model detects the spikes and the overall behaviour of the changes in the concentrations of the examined pollutant PM10 and outperforms the SARIMA forecasts.

Figure 10 presents the aggregated average daily forecasts of the SARIMA model S1 and GPS model R5 compared with the actual PM10 mean values for a 5-days-ahead horizon. It is observed very good agreement of the predictions with the measured data.

Figure 10 Average daily values, forecasted using the models S1 and R5 compared with the actual PM10 data (see online version for colours)



5 Conclusions

This paper presented a statistical investigation of the main air pollutant PM10 in relation to six meteorological factors, describing ambient air quality in the town of Kardzhali, in the southern region of Bulgaria. The examined data are characterised by high variability and constantly exceed the permissible limits for PM10, indicating that the status of this ecological indicator is troubling.

Two different approaches were applied to investigate the PM10 time series – stochastic SARIMA and the new GPS method. Firstly, it was found that the time series had no trend and showed a 24-hour cycle. When constructing SARIMA models, time variables *DATE_*, *DAY_*, *HOUR_* and seasonality $s = 24$ were defined. In order to improve distribution and stabilise the variance a preliminary transformation of PM10 variable was performed. Three models of the type SARIMA(1, 0, 7)(2, 0, 1)$_{24}$ were built and analysed with and without the use of the six meteorological variables. The best SARIMA model S1 takes into account three meteorological variables - wind speed, air temperature and pressure, fits the data with a high coefficient of determination $R^2 = 90\%$ and has a good forecasting ability. We will point out that the model S3, built without meteorological predictors, also shows good statistics (see Table 2). This can be explained by the fact that the autoregressive term of the previous hour implicitly includes the influence of all observed and unobserved factors on the pollutant values. The specificity of modelling with SARIMA in this study was the determination of the large number of parameters for the meteorological time series to find a valid model that requires sophisticated efforts.

The pollutant was also modelled with the new GPS method, supported by three data mining techniques - gradient boosting trees (TreeNet), ISLE and RuleLearner. Three time predictors were entered – *DAYS_HOURS*, *DAY* and *HOUR* to describe the data as a time series. In the construction of all models, before applying GPS regularised regression, observations were classified by the TreeNet, depending on similar weather conditions. This way, a huge amount of new predictors using these classification trees are determined. On this basis, GPS models have been obtained with all of these secondary predictors, as well as with all and with the most important meteorological variables as raw predictors. GPS models with high fit and forecasting quality have been built. The best GPS model R5 explains up to R^2 = 82% of the data. The relative importance of the raw predictors in the models was established, indicated that a weak wind, air temperature, air pressure and relative humidity are more important. According to the obtained results, the GPS models showed very good accuracy both in prediction and forecasts in comparison to the known data.

The developed empirical models by means of SARIMA and GPS methods have been applied for short-term forecasting for a period of 5-days-ahead using data non-involved in the construction of models and the results demonstrated very good performance. The forecasting results of the GPS models outperform the results of the SARIMA models.

The obtained results showed that a common basic feature of the GPS approach, enhanced by data mining techniques, is the ability to build flexible, adequate, high-performance models with very good forecasting capabilities. The method is comparable to the well-known classical Box-Jenkins methodology for time series analysis. One disadvantage is the final model complexity that leads to some difficulties in the direct interpretation of the results.

Acknowledgements

This work has been accomplished with the financial support by the Grant No BG05M2OP001-1.001-0003, financed by the Science and Education for Smart Growth Operational Program (2014-2020) and co-financed by the European Union through the European structural and Investment funds.

References

Air Quality Standards (2013) *Environment*, European Commission [online] http://ec.europa.eu/environment/air/quality/standards.htm (accessed April 2018).

ALADIN (2018) *Project for Weather Forecasts*, Bulgaria [online] http://www.weather.bg/0index.php?koiFail=cities1&lng=1&ci=Kardzhali&gr=Kardzhali (accessed April 2018).

Al-Madfai, H., Geens, A.J. and Snelson, D.G. (2010) 'Modelling the multi-year maximum daily PM10 concentration in Edinburgh: an application of the variability decomposition transfer function model', in Brebbia, C.A. and Longhurst, J.W.S. (Eds.): *Air pollution XVIII*, WIT Tran Ecol Environ, WIT Press, Vol. 136, pp.349–356, DOI: 10.2495/AIR100311.

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G. and Di Carlo, P. (2017) 'Recursive neural network model for analysis and forecast of PM10 and PM2.5', *Atmospheric Pollution Research*, Vol. 8, No. 4, pp.652–659.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.S. (1994) *Time Series Analysis, Forecasting and Control*, 3rd ed., Prentice-Hall, New Jersey.

Davalos, A.D., Luben, T.J., Herring, A.H. and Sacks, J.D. (2017) 'Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures', *Annals Of Epidemiology*, Vol. 27, No. 2, pp.145–153, e1.

Directives on Ambient Air (AAQ) (2008) *Quality Assessment and Management*, Official Journal of the European Union, L 152/1 [online] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:152:0001:0044:EN:PDF (accessed April 2018).

Dockery, D.W. and Pope, C.A. (1994) 'Acute respiratory effects of particulate air pollution', *Annual Review of Public Health*, Vol. 15, pp.107–132, DOI: 10.1146/annurev.pu.15.050194.000543.

Dotse, S.Q., Petra, M.I., Dagar, L. and De Silva, L.C. (2018) 'Application of computational intelligence techniques to forecast daily PM10 exceedances in Brunei Darussalam', *Atmospheric Pollution Research*, Vol. 9, No. 2, pp.358–368.

European Environment Agency (EEA, 2017) *Air quality in Europe – 2017* [online] https://www.eea.europa.eu/publications/air-quality-in-europe-2017 (accessed April 2018).

Executive Environment Agency (ExEA, 2018) Bulgaria [online] http://pdbase.government.bg/airq/bulletin-en.jsp (accessed April 2018).

Friedman, J.H. (2001) 'Greedy function approximation: A gradient boosting machine'*, 1999 Reitz Lecture, Annals of Statistics*, Vol. 29, No. 5, pp.1189–1232.

Friedman, J.H. (2012) 'Fast sparse regression and classification', *International Journal of Forecasting*, Vol. 28, No. 3, pp.722–738.

Friedman, J.H. and Popescu, B.E. (2003) *Importance Sampled Learning Ensembles*, Technical Report, Stanford University, Department of Statistics [online] http://www-stat.stanford.edu/~jhf/ftp/isle.pdf (accessed April 2018).

Friedman, J.H. and Popescu, B.E. (2005) *Predictive Learning via Rule Ensembles*, [online] http://www-stat.stanford.edu/~jhf/ftp/RuleFit.pdf (accessed April 2018).

Ganesh, S.S., Arulmozhivarman, P. and Rao, Tatavarti (2017) 'Forecasting air quality index using an ensemble of artificial neural networks and regression models', *Journal of Intelligent Systems*, Vol. 34, No. 1, pp.1–11 [online] https://doi.org/10.1515/jisys-2017-0277.

Grange, S.K., Carslaw, D.C., Lewis, A.C., Boleti, E. and Hueglin. C. (2018) 'Random forest meteorological normalisation models for Swiss PM10 trend analysis', *Atmospheric Chemistry and Physics, Discussions*, DOI: 10.5194/acp-2017-1092.

Hoerl, A.E. and Kennard, R.W. (1970) 'Ridge regression: biased estimation for nonorthogonal problems', *Technometrics*, Vol. 42, No. 1, pp.80–86, DOI: 10.1080/00401706.1970.10488634.

Ivanov, A.V. and Gocheva-Ilieva, S.G. (2013) 'Short-time particulate matter PM10 forecasts using predictive modeling techniques', *AIP Conference Proceedings*, Vol. 1561, pp.209–218, DOI: 10.1063/1.4827230.

Lee, N.U., Shim, J.S., Ju, Y.W. and Park, S.C. (2017) 'Design and implementation of the SARIMA–SVM time series analysis algorithm for the improvement of atmospheric environment forecast accuracy', *Soft Computing*, Vol. 21, DOI: 10.1007/s00500-017-2825-y.

Lima, E.A.P., Guimaraes, E.C., Pozza, S.A., Barrozo, M.A.S. and Coury J.R. (2009) 'A study of atmospheric particulate matter in a city of the central region of Brazil using time-series analysis', *International Journal of Environmental Engineering*, Vol. 1, No. 1, pp.80–94.

Liu, P.W.G. (2009) 'Simulation of the daily average PM10 concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis', *Atmospheric Environment*, Vol. 43, pp.2104–2113, DOI: 10.1016/j.atmosenv.2009.01.055.

Moazami, S., Noori, R., Amiri, B.J., Yeganeh, B., Partani, S. and Safavi, S. (2016) 'Reliable prediction of carbon monoxide using developed support vector machine', Atmospheric *Pollution Research*, Vol. 7, No. 3, pp.412–418, DOI: 10.1016/j.apr.2015.10.022.

Nisbet, R., Elder, J. and Miner, G. (2009) *Handbook of Statistical Analysis and Data Mining Applications*, Academic Press Elsevier, Burlington, MA.

Noori, R., Hoshyaripour, G., Ashrafi, K. and Araabi, B.N. (2010) 'Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration', *Atmospheric Environment*, Vol. 44, No. 4, pp.476–482, DOI: 10.1016/ j.atmosenv.2009.11.005.

Salford Predictive Modeler Users Guide (2013) Introducing Generalized PathSeeker, Salford Systems, San Diego [online] http://media.salford-systems.com/pdf/spm7/IntroGPS.pdf (accessed April 2018).

Shahraiyni, T.H. and Sodoudi, S. (2016) 'Statistical modeling approaches for PM10 prediction in urban areas: a review of 21st-century studies', *Atmosphere*, Vol. 7, No. 2 [online] http://www.mdpi.com/2073-4433/7/2/15/htm (accessed April 2018).

Siwek, K., Osowski S. and Sowinski, M. (2011) 'Evolving the ensemble of predictors model for forecasting the daily average PM10', *International Journal of Environment and Pollution*, Vol. 46, Nos. 3/4, pp.199–215.

SPM v8.2 (Salford Predictive Modeler software suite) (2017), [online] http://www.salford-systems.com/products/spm, (accessed 1 April 2018).

Stoimenova, M., Voynikova, D., Ivanov, A., Gocheva-Ilieva, S. and Iliev, I. (2017) 'Regression trees modeling and forecasting of PM10 air pollution in urban areas', *AIP Conference Proceedings*, Vol. 1895, No. 030005, pp.1–10, DOI: 10.1063/1.5007364.

Tabachnik, B.G. and Fidell, L.S. (2007) *Using Multivariate Statistics*, 5th ed., Allyn and Bacon/Pearson Education, Boston, MA.

Tibshirani, R. (1996) 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society*, Ser. B, Vol. 58, No. 1, pp.267–288.

Ul-Saufie, A.Z., Yahaya, A.S., Ramli, N.A., Rosaida, N. and Hamid, H.A. (2013) 'Future daily PM10 concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA)', *Atmospheric Environment*, Vol. 77, pp.621–630, DOI: 10.1016/j.atmosenv.2013.05.017.

Wang, P., Liu, Y., Qin, Z. and Zhang, G. (2015) 'A novel hybrid forecasting model for PM10 and SO2 daily concentrations', *Science of the Total Environment*, 1 February, Vol. 505, pp.1202–1212.

Whalley, J. and Zandi, S. (2016) 'Particulate matter sampling techniques and data modelling methods', in Sallis P. (Ed.): *Air Quality – Measurement and Modeling*, INTECH, Vol. 2, pp.29–54.

Wilks, D.S. (2011*) Statistical Methods in the Atmospheric Sciences*, 3rd ed., Elsevier, Amsterdam.

Yeo, I.K. and Johnson, R.A. (2000) 'A new family of power transformations to improve normality or symmetry', *Biometrika*, Vol. 87, No. 4, pp.954–959, DOI: 10.1093/biomet/87.4.954.

Zheleva, I., Veleva, E. and Filipova, M. (2017) 'Analysis and modeling of daily air pollutants in the city of Ruse, Bulgaria', in Todorov, M. (Ed.): *AIP Conference Proceedings*, Vol. 1895, No. 030007, DOI:10.1063/1.5007366.

Zou, H. and Hastie, T. (2005) 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society*, Vol. 67, Ser. B, pp.301–320.