Time series analysis and forecasting for air pollution in small urban area: An SARIMA and factor analysis approach

Article *in* Stochastic Environmental Research and Risk Assessment · May 2013

DOI: 10.1007/s00477-013-0800-4

4 authors:

Snezhana Gocheva-Ilieva
University of Plovdiv "Paisii Hilendarski"
93 PUBLICATIONS **264** CITATIONS

SEE PROFILE

A. Ivanov
Plovdiv University "Paisii Hilendarski"
21 PUBLICATIONS **61** CITATIONS

SEE PROFILE

Desislava Stoyanova Voynikova
University of Plovdiv Paisii Hilendarski
24 PUBLICATIONS **70** CITATIONS

SEE PROFILE

Doychin Boyadzhiev
Plovdiv University "Paisii Hilendarski"
33 PUBLICATIONS **329** CITATIONS

SEE PROFILE

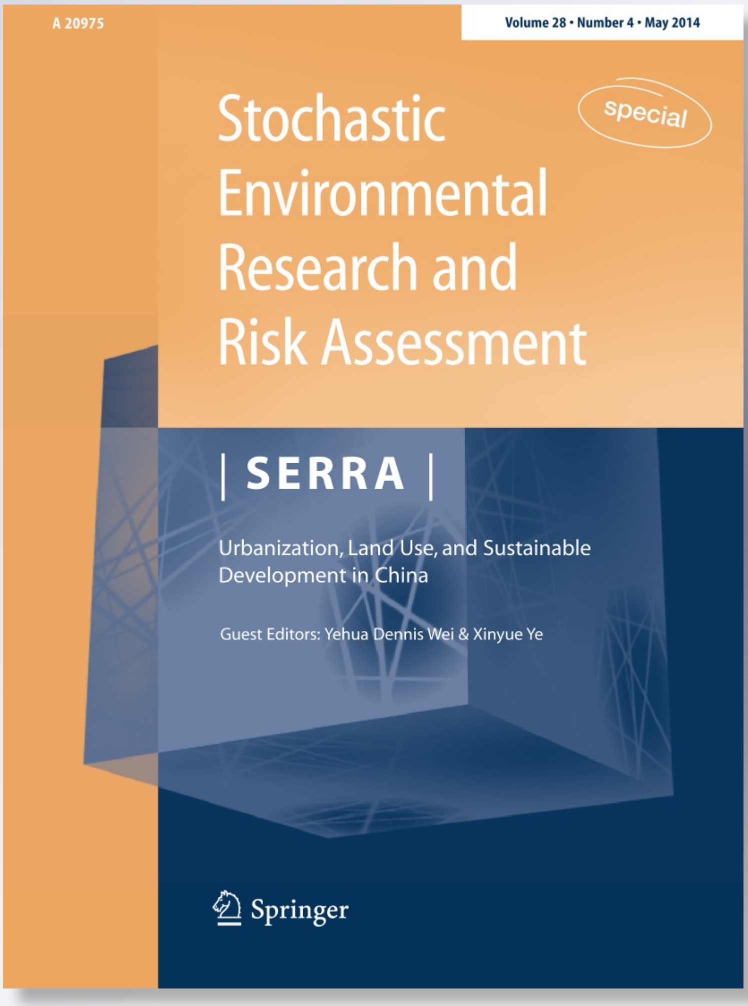
Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach

Snezhana Georgieva Gocheva-Ilieva, Atanas Valev Ivanov, Desislava Stoyanova Voynikova & Doychin Todorov Boyadzhiev

Stochastic Environmental Research and Risk Assessment

ISSN 1436-3240
Volume 28
Number 4

Stoch Environ Res Risk Assess (2014)
28:1045-1060
DOI 10.1007/s00477-013-0800-4


A 20975
Volume 28 • Number 4 • May 2014
Stochastic Environmental Research and Risk Assessment
special
| SERRA |
Urbanization, Land Use, and Sustainable Development in China
Guest Editors: Yehua Dennis Wei & Xinyue Ye
Springer

Stoch Environ Res Risk Assess (2014) 28:1045–1060
DOI 10.1007/s00477-013-0800-4

ORIGINAL PAPER

Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach

Snezhana Georgieva Gocheva-Ilieva · Atanas Valev Ivanov · Desislava Stoyanova Voynikova · Doychin Todorov Boyadzhiev

Published online: 25 September 2013

Abstract Despite the existing public and government measures for monitoring and control of air quality in Bulgaria, in many regions, including typical and most numerous small towns, air quality is not satisfactory. In this paper, factor analysis and Box–Jenkins methodology are applied to examine concentrations of primary air pollutants such as NO, NO_2, NO_x, PM10, SO_2 and ground level O_3 in the town of Blagoevgrad, Bulgaria within a 1 year period from 1st September 2011 to 31st August 2012, based on hourly measurements. By using factor analysis with PCA and Promax rotation, a high multicollinearity between the six pollutants has been detected. The pollutants were grouped in three factors and the degree of contribution of the factors to the overall pollution was determined. This was interpreted as the presence of common sources of pollution. The main part of the study involves the performance of time series analysis and the development of univariate stochastic seasonal autoregressive integrated moving average (ARIMA) models with recording on a hourly basis as seasonality. The study also incorporates the Yeo–Johnson power transformation for variance stabilizing of the data and model selection by using Bayersian information criterion. The obtained SARIMA models demonstrated very good fitting performance with regard to the observed air pollutants and short-term predictions for 72 h ahead, in particular in the case of ozone and particulate matter PM10. The presented statistical approaches allow the building of non-complex models, effective for short-term air pollution forecasting and useful for advance warning purposes in urban areas.

S. G. Gocheva-Ilieva (✉) · A. V. Ivanov ·
D. S. Voynikova · D. T. Boyadzhiev
Department of Applied Mathematics and Modeling, Plovdiv University Paisii Hilendarski, 24 Tzar Assen Street,
4000 Plovdiv, Bulgaria
e-mail: snegocheva@yahoo.com; snow@uni-plovdiv.bg

A. V. Ivanov
e-mail: aivanov_99@yahoo.com

D. S. Voynikova
e-mail: desi_sl2000@yahoo.com

D. T. Boyadzhiev
e-mail: dtb@uni-plovdiv.bg

Keywords Air quality modeling · Air pollution forecast · Factor analysis · Time series · SARIMA · Seasonal Box–Jenkins models · Univariate stochastic models

Mathematics Subject Classification 62M10 · 62M20 · 62P12

1 Introduction

Nowadays, the continuous and strict monitoring and forecasting of ambient air pollutants is of great importance in the process of evaluating regulatory control measures related to air quality. Many countries increasingly include active monitoring and control of key pollution indicators within all regions of their territory. In Bulgaria, 12 types of pollutants are systematically monitored by more than 36 automated stations run by the Executive Environment Agency which manages and coordinates activities related to the control and environmental protection of the country. The European and national prescribed pollution levels and limits are also monitored (EEA 2012, 2013; Directive 2008; Air Quality Standards 2013). Atmospheric air quality reports for the various regions of the country are regularly published. As a result, huge amount of data are accumulated. This allows for the performance of various analyses, including statistical ones, in order to find general patterns and dependencies for different time periods and relationships between observed air characteristics.

This is directly related to development and application of suitable tools for data analysis and forecasting. In particular, the classical techniques of PCA and factor analysis are important statistical instruments frequently used in environmental sciences. The main advantages of these methods are that they reveal strong correlation relationships between observed variables and allow their grouping in new variables (factors) in order to reduce the dimensions of the complex data structure (Jolliffe 2002; Kim and Mueller 1986). The factors can subsequently be used to build regression or other type of models. Typical examples of such models are presented in (Blifford and Meeker 1967; Henry and Hidy 1979; Lengyel et al. 2004; Huang et al. 2011). More applications of factor analysis in ecology are given in (Kaplunovsky 2005).

Other parametric methods widely used for times series analysis and forecasting are the autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) models, known also as Box–Jenkins stochastic models (Box and Jenkins 1976). Some of the main advantages of the Box–Jenkins approach are (Box and Jenkins 1976; Pankratz 1983; Chatfield 1996, 2000; McBerthouex and Brown 2002): (i) Its applicability for modeling and forecasting practically any time series, which is stationary or can be reduced to stationary via a differencing procedure; (ii) The ability to extract all the trends and serial correlations in the data with a minimized sequence of white noise (shock) through inclusion in one general model equation that gets to the basis of historical data development; (iii) The method has been incorporated into many standard software packages such as SPSS, Statistica, *R* and many others (Comparison of Statistical Packages 2013), which speeds up and facilitates the modeling process significantly. As disadvantages of the method we can note that since Box–Jenkins models are empirical, an identification-estimation-diagnosis procedure must always be carried out. Also, for time series analysis, at least an additional 50–100 observations are needed (Box and Jenkins 1976; Pankratz 1983, Milionis and Davies 1994). This might present a problem, for example with yearly data.

In principle, the processes of air pollution in the atmosphere are strongly governed by meteorology (e.g., see Jacobson 2005). However, in so called univariate models, it is assumed that the final concentration of air pollutants in the atmosphere is the final result of all the complex interplay of meteorology, chemistry, transport, diffusion etc. Therefore, the combined information of their effect on air pollutant concentration is contained in the corresponding time series in a stochastic way. With this approach calculations are simplified and performed only using the time series of the pollutant without explicit inclusion of meteorological or other measurements. Moreover, some analytics argue that univariate Box–Jenkins models frequently approach or exceed the forecasting accuracy of multiple-series models, especially for short-term forecasts (Pankratz 1983).

Box–Jenkins methodology is widely applied in air quality studies. Taking into account the influence of meteorological factors, ARIMA models have been built to predict submicron particle concentrations (Jian et al. 2012), daily average PM10 concentrations (Liu 2009), ozone concentration at urban and rural areas (Dueñas et al. 2005). In (Slini et al. 2002) univariate ARIMA models are obtained for maximum ozone concentration forecasts in using 9-year air quality observations. Good index of agreement, accompanied by a weakness in forecasting alarms are reported. A number of univariate ARIMA/SARIMA models are developed in (Sharma et al. 2009) in order to analyze and forecast monthly maximums of the 24-h average time-series data for SO_2, NO_2 and suspended particulate matter concentration in an urban area. In (Kumar and Jain 2010) univariate stochastic ARIMA models are developed to forecast daily mean ambient air pollutants O_3, CO, NO and NO_2 concentration at an urban traffic site. During the selection of the best models, comparisons have been made utilizing various data transformations and information criteria.

This paper presents a statistical study of air pollution in the town of Blagoevgrad, which is a typical medium town in South-West Bulgaria, situated within a valley. It can be mentioned here, the officially available air pollution statistics and planned activities for improvement of particulate matter PM10 can be found in the Program for reduction of harmful emissions in atmospheric air (PM10) on the territory of Blagoevgrad (Program 2011), for the period between 1st January 2008 and 1st March 2011. This document offers a detailed analysis of specific sources of pollution (domestic heating, factories, road traffic, etc.). There are no other studies published on this topic for this region. We have to note the recently conducted similar studies for the town of Burgas and the town of Shumen, Bulgaria (Petelin et al. 2013; Ivanov et al. 2012).

The main purpose of this study is to establish the dependence of variation in the levels of air pollution and possible combined effects within 1 year, and to reveal the sources of these. Two statistical approaches—factor analysis and SARIMA are applied to describe the actual environmental status, as well as to find out the more appropriate forecast methods for this type of data. The two approaches are complementary and clarify the various aspects of the behavior of air pollutants.

The questions considered in the study are: (i) Identifying correlation type dependences and grouping of observed air pollutants using the method of factor analysis to explain mutual effects of pollution; (ii) Conducting time series analysis by determining seasonal ARIMA (based on hourly data) relevant parametric models of pollutants; (iii) Analysis and diagnostics of constructed models; (iv) Application of models for short-term forecasting; (v) Interpretation of the results and definition of the conditions contributing to the exceeding of national and European concentration norms for the considered air pollutants.

The study was carried out by using IBM SPSS 19 software package for Windows (SPSS 2013) and EViews 7 for Windows (EViews 2013).

2 Data description

2.1 Study area

We will examine air quality in the town of Blagoevgrad, which is a typical representative of a small urban region. The town is located in Southwest Bulgaria in valley of the Struma river, 100 km away from the capital city of Sofia. The exact coordinates are: 42°01′N, 23°06′E, altitude of 360 m (Blagoevgrad 2013). The town is characterized by mountainous and valley relief with plenty of vegetation—parks and forests. The climate is transitional continental with a strong Mediterranean influence. Its population is around 70,000 people with a tendency to increase. There is little road traffic as it is located away from busy highways. The buildings are typically low- to medium-size. There is no immediate pollution from other nearby towns or cities.

2.2 Data

We examine data for six of the main air pollutants in the town of Blagoevgrad during a 1 year period from 1st September 2011 to 31st August 2012. The observed pollutants are concentrations of particulate matter (PM10), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), nitrogen oxide (NO), nitrogen oxides (NO_x), and ground level ozone (O_3). The data are expressed in units of mass concentration of pollutants in $\mu g/m^3$, except for NO_x—in ppb. Here, NO_x includes pollution from all kinds of nitrous oxides.

The measurements and the data used have been collected and processed by the monitoring station of ExEA, through unified methods, accredited under the BS EN ISO/IEC 17025—General requirements for competence in testing and calibration from EA BAS (ISO 2013; National System 2013).

The total volume of data is taken from 8,744 cases, by hours. There are no missing data values. It can be noted, that in this study data are modeled without removing outliers. This is done in view of the possibility of comparing the results with further nonparametric statistical analysis.

It also needs to be noted that when comparing the data between 2008 and 2012, the types of pollutants in the town of Blagoevgrad demonstrate similar behavior within 1 year. For this reason and in order to simplify calculations, the last year has been chosen in order to record the most recent data. Moreover, the goal of the study is to demonstrate the capabilities of the presented approaches, which can also be applied to other observation sets, including for shorter or longer periods of time.

The corresponding plots of the observed six pollutants are shown in Figs. 1, 2, 3, 4, 5, and 6 in blue color.

Table 1 provides brief descriptive statistics of the data: total number, minimum, maximum, average, standard deviation, coefficients of skewness and kurtosis. The official Bulgarian national norms of admissible and allowed concentrations of the examined air pollutants are given. The indices of skewness and kurtosis are usually used to check the properties of symmetry and flatness of the function of the density and data distribution in the time series. The coefficient of skewness is quite sensitive to extremes and discontinuous fluctuations. Kurtosis is an indicator for the presence of interruptions in the time series. Table 1 indicates that the highest value of the coefficient of skewness is that of NO, which corresponds to sharp increases in the data as presented in Fig. 2. The coefficient of kurtosis also demonstrates a very high value for NO, as well as for SO_2, which corresponds to the existing discontinuities in the data in Figs. 2 and 5, respectively.

Because of the hourly nature of the data, the seasonality in the time series in our models can be considered by hours, so that it will affect the short-term forecasting result.

The more detailed analysis of the results in Table 1 shows that for some variables (O_3, PM10) the corresponding standard deviations and mean values are almost equal. Also, for other pollutants, the standard deviation is about two times bigger than the respective means. This indicates that the sensitivity to uncertainties for these pollutants is high. Application of special approaches, such as in (Dimov et al. 2010, 2013), might help understand further the detailed impact of data properties on the obtained parametric models.

As it is well known, the application of parametric models requires normal or near to normal distribution of variables. The last column in Table 1 shows the results of a Kolmogorov–Smirnov sample test for normality, carried out by SPSS (SPSS IBM Statistics 2013). The obtained K–S statistics indicate the non-normality of the data, except for O_3. For improving the distribution and minimizing the variability of the data, different transformations could be applied prior to constructing models.

To this reason we use the following Yeo–Johnson power transformation (Yeo and Johnson 2000), which represent an improvement over the Box–Cox transformation family

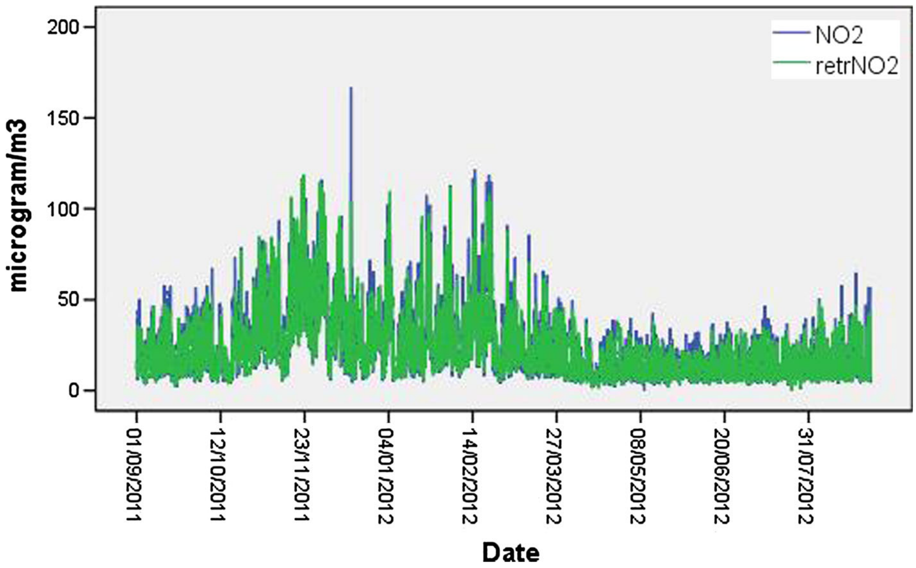


Fig. 1 Observed values for NO_2 (*in blue*) and predicted values (*in green*). (Color figure online)

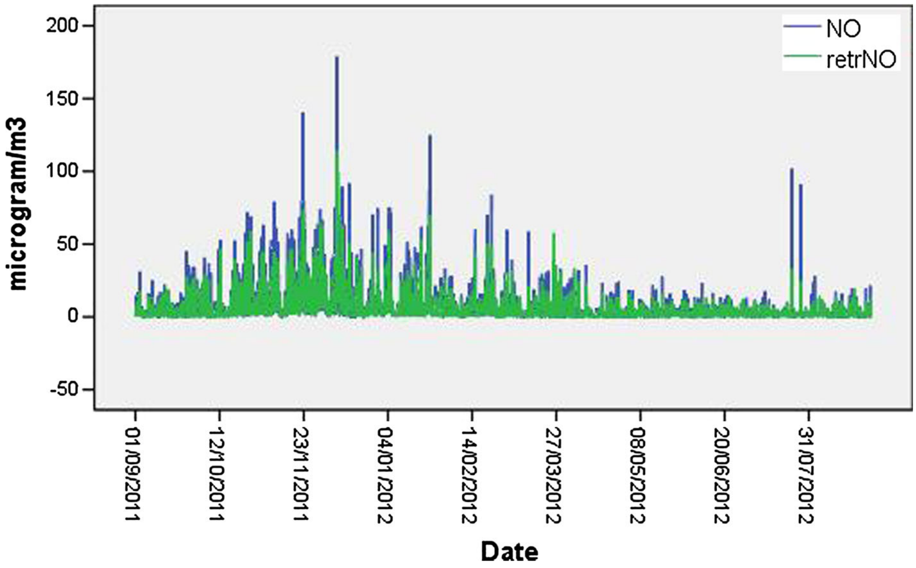


Fig. 2 Observed values for NO (*in blue*) and predicted values (*in green*). (Color figure online)

(Box and Cox 1964) and is appropriate for an arbitrary sign of data:

$$\psi_{YJ}(\lambda,x)=\begin{cases}\left\{(x+1)^{\lambda}-1\right\}/\lambda & x\geq 0,\lambda\neq 0\\ \log(x+1) & x\geq 0,\lambda=0\\ -\left\{(-x+1)^{2-\lambda}-1\right\}/(2-\lambda) & x<0,\lambda\neq 2\\ -\log(-x+1) & x<0,\lambda=2\end{cases}$$
$$\lambda\in[-2,2] \tag{1}$$

For our data, the Yeo–Johnson transformation coefficients for any of the observed variables were found using simple procedure of attempts from the sequence [−2, −1.9, −1.8,…,2]. The obtained coefficients λ and descriptive statistics for calculated transformed data are shown in Table 2. It can be seen that the transformed data variables satisfy the Kolmogorov–Smirnov test of normality at 0.05 level of significance and can be considered to be normally distributed.

3 Factor analysis approach

Factor analysis is widely applied methodology in atmospheric science for processing of time series data. The goal of this approach is to establish the presence or absence of combined interactions between the investigated pollutants. The presence of such interactions is to be interpreted as resulting from common sources of pollution.

The statistical technique of factor analysis allows for the reduction of the number of mutually dependent variables

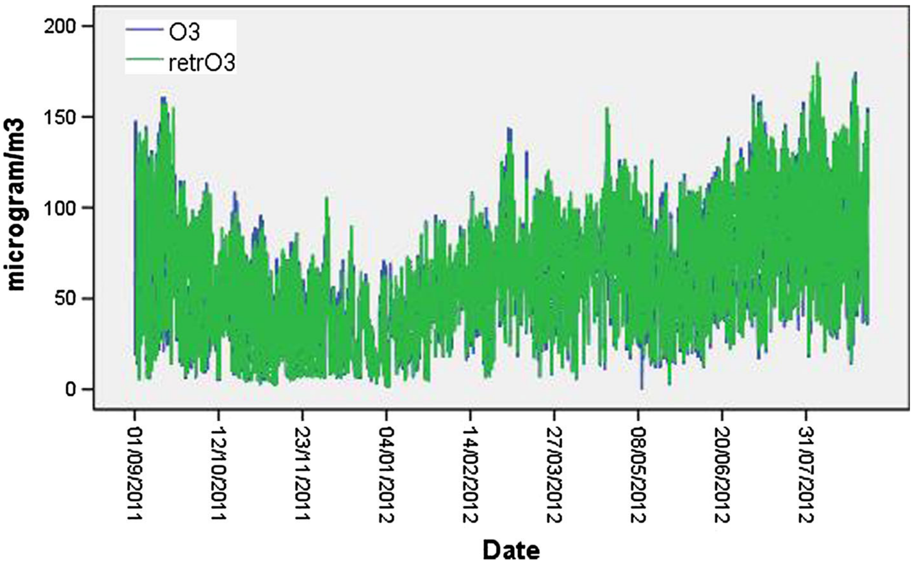


Fig. 3 Observed values for O_3 (*in blue*) and predicted values (*in green*). (Color figure online)

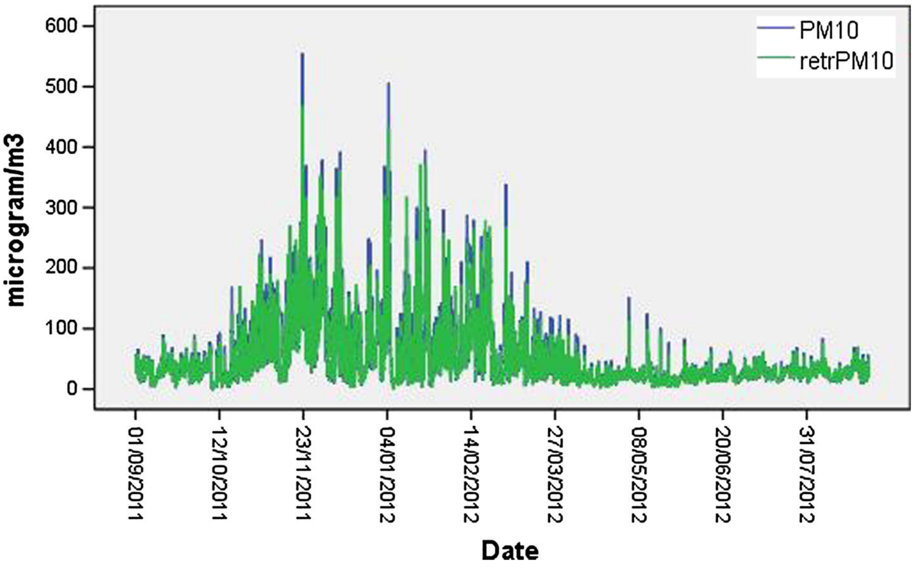


Fig. 4 Observed values for PM10 (*in blue*) and predicted values (*in green*). (Color figure online)

by grouping together strongly correlated variables. The method is also used for clarifying similarities and differences in a multidimensional dataset (Jolliffe 2002; Kim and Mueller 1986). The presence of high coefficients (over 0.5) in the correlation matrix of the data often indicates that the correlation matrix is singular (i.e. its determinant is close to 0). When two or more variables correlate strongly, in factor analysis, these are represented by a general latent (artificial) variable, called a factor. During this process, some of the information is lost but the relationships between the grouped variables are found. Factor analysis employs various methods for extracting and transforming (rotating) the factors. Another key point is determining the number of factors which will replace the initial independent variables. The choice of a specific method, transformation, and number of factors is up to the researcher.

Factor analysis was applied to the six air pollutants under investigation. The procedure comprises five steps: (a) calculation of the correlation matrix, (b) testing the adequacy of factor analysis, (c) factor extraction, (d) factor rotation and (e) score calculation of factor variables. The following results were obtained.

The corresponding correlation matrix is given in Table 3. It shows that the determinant is small enough (8.53×10^{-7}) and there are large statistically significant correlation coefficients. It can be added that O_3 has negative correlations with all other observed variables. This means that the behavior of O_3 is inversely proportional to all other pollutants.

Generally, in factor analysis, the measures of adequacy are the statistical indices of Kaiser–Mayer–Olkin (KMO) measure of sampling adequacy and Bartlett's test of

Fig. 5 Observed values for SO_2 (*in blue*) and predicted values (*in green*). (Color figure online)

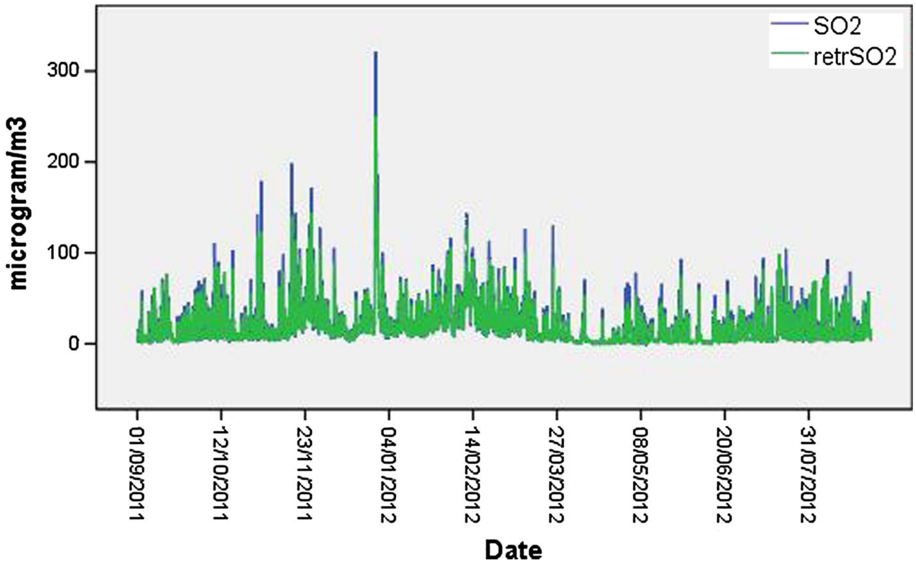


Fig. 6 Observed values for NO_x (*in blue*) and predicted values (*in green*). (Color figure online)

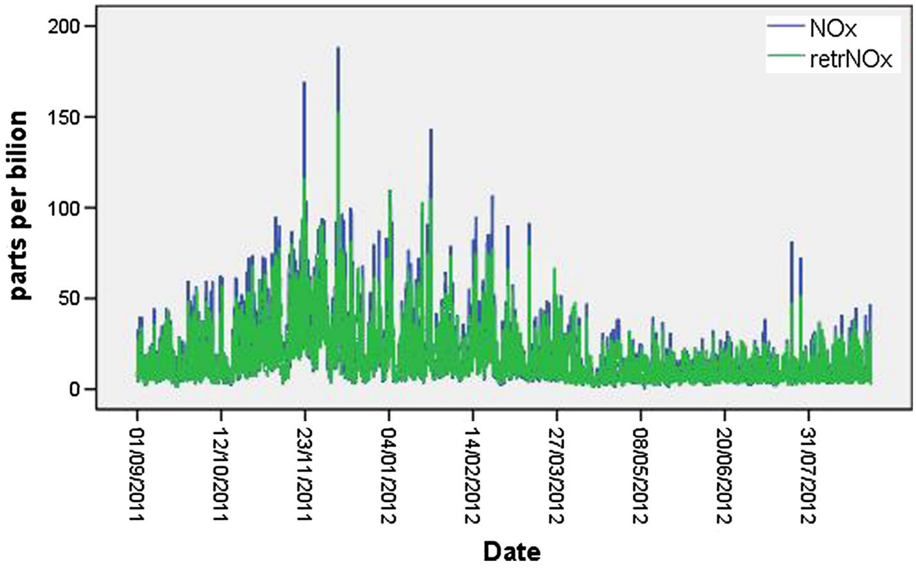


Table 1 Descriptive statistics of observed air pollutants of town of Blagoevgrad

Variable	Threshold limit	Official hour limit norm	Minimum	Maximum	Mean	Std. Dev.	Skewness	Kurtosis	K–S test
NO_2, μg/m^3	40[a]	200	0.121	166.6	21.70	27.07	1.90	4.79	0.137
NO, μg/m^3	10[a]	–	0.000	179.1	5.75	10.50	5.14	41.02	0.293
O_3, μg/m^3	120[b]	180	1.078	174.5	61.90	64.52	0.37	−0.53	0.043
PM10, μg/m^3	40[a]	50	0.467	555.5	46.87	38.90	2.87	11.77	0.213
SO_2, μg/m^3	125[b]	350	0.000	321.5	17.44	25.33	3.60	21.83	0.212
NO_x, ppb	30[a]	–	1.184	188.2	16.00	24.64	2.81	12.27	0.188

Std. Error of the Skewness is 0.026, Std. Error of the Kurtosis is 0.052

[a] Average per year

[b] Average per day

sphericity, where KMO needs to be more than 0.5, and Bartlett's test of sphericity to be statistically significant (Sig < 0.05). In our case KMO $= 0.627$ and Bartlett's test is significant at Sig. $= 0$, which shows that there is a relationship between most of variables, and that the performance of factor analysis is recommended. Only SO_2 does not have

Table 2 Descriptive statistics of transformed data of town of Blagoevgrad

Transformed variable	Coefficient λ	Minimum	Maximum	Mean	Std. Dev.	Skewness	Kurtosis	K–S test
$trNO_2$	0	−2.12	5.12	2.82	0.73	0.01	−0.18	0.02
trNO	−0.2	−9.80	3.23	0.75	0.97	−0.28	1.58	0.02
trO_3	0.8	−1.25	76.44	31.72	15.33	0.14	−0.68	0.03
trPM10	−0.2	−0.82	3.59	2.47	0.42	−0.46	1.58	0.03
$trSO_2$	0	−5.95	5.78	2.30	1.09	−0.17	0.04	0.03
$trNO_x$	−0.2	0.17	3.25	1.89	0.48	−0.04	−0.34	0.02

Std. Error of the Skewness is 0.026, Std. Error of the Kurtosis is 0.052

Table 3 Correlation matrix of six air pollutants (non-transformed)

	NO_2 µg/m^3	NO µg/m^3	O_3 µg/m^3	PM10 µg/m^3	SO_2 µg/m^3	NO_x ppb
Correlation						
NO_2	1	0.63	−0.56	0.84	0.35	0.91
NO	0.63	1	−0.43	0.63	0.15	0.90
O_3	−0.56	−0.41	1	−0.45	−0.07	−0.54
PM10	0.84	0.629	−0.45	1	0.33	0.82
SO_2	0.35	0.15	−0.07	0.33	1	0.28
NO_x	0.91	0.90	−0.54	0.82	0.28	1

Determinant = 8.53E−007. Significance levels of all correlation coefficients are 0.000

Table 4 Pattern matrix from factor analysis

Variables	Component		
	Factor *F*1	Factor *F*2	Factor *F*3
NO_2, µg/m^3	1.075		
NO, µg/m^3	0.979		
NO_x, ppb	0.775		
PM10, µg/m^3	0.703		
O_3, µg/m^3		−1.027	
SO_2, µg/m^3			1.014

Extraction method: principal component analysis

Rotation method: Promax with Kaiser normalization; rotation converged in 5 iterations

Factor loadings less than 0.5 are omitted

high correlation coefficients and can be considered as a unique variable. Table 3 shows, for example, that the strongest correlations are these between NO_2 and NO_x (0.91), and NO and NO_x (0.90), PM10 and NO_2 (0.84) and PM10 and NO_x (0.82), PM10 and NO (0.629), which are expected to be grouped in one factor. All correlation coefficients are statistically significant with Sig. = 0.

The factors have been extracted using the PCA method (Jolliffe 2002), resulting in three factors. The Promax method of factor rotation turned out to be the most appropriate one. For our data, it provides a sharp distinction between factors as compared to Varimax and other popular methods of rotation. The Promax rotation is a well-known technique in ecology and climatology (Richman 1986). For our data, it is used to enter the three factors in a non-orthogonal coordinate system. The resulting rotated factor loadings are given in Table 4, where loadings under 0.5 have been ignored.

The pollutants are clearly divided into three groups. The requirement of factor analysis that a given variable should only participate in one factor has been met. The resulting factors are as follows:

$$F1 = \{NO_2,\ NO,\ NO_x,\ PM10\}, F2 = \{O_3\},\\ F3 = \{SO_2\} \tag{2}$$

These three factors account for 90.74 % of the total variance of the data. The minimum recommended portion of dataset is 80 % (Jolliffe 2002). The partial contribution of the factors is respectively: 45 % for *F*1, 27 % for *F*2, 19 % for *F*3. The separation of the two pollutants O_3 and SO_2 as single factors is according to the expectations, considering the correlation matrix from Table 3.

These three groups have been identified in more detail further on in the section interpreting the obtained results.

4 ARIMA and SARIMA approach

ARIMA and SARIMA are widely used general classes of models, introduced by Box and Jenkins in 1970 (Box and Jenkins 1976).

Current publications mainly include studies where stochastic ARIMA and SARIMA models are built based on daily mean observations. In practice, the changes in air pollutant concentrations are usually observed within shorter time intervals (of several hours) and their values at almost the same times during the previous few days can be taken into account. For this reason, in the face of such cyclic recurrence, we consider SARIMA more adequate as it would allow the development of more accurate models. Respectively, with hourly data, forecasting pollution concentrations will be more accurate but within several days.

We will add that in cases with highly variable data, more powerful methods may also be used, including nonparametric and hybrid ones, such as neural networks, tree-based and others (Gardner and Dorling 1999; Brunelli et al. 2007; Díaz-Robles et al. 2008; Kim 2010; Kim and Kumar 2005, Polydoras et al. 1998). However, the requirement for these methods is the use of meteorological, transport, and other data.

The general form of ARIMA (*p*, *d*, *q*) includes the following nonnegative integer general parameters: *p* is the number of the parameters describing the autoregressive process (AR), *d* is the number of parameters for trend process (I) and *q* is the number of parameters for moving average process (MA). Usually, the estimation of the parameters is obtained by an iterative procedure of minimizing sum of squares, as with non-linear regression.

The time values are denoted by $t = 1, 2, 3,\ldots,n$, where *n* is the total number of observations in the time series. X_t denotes the value of a time series variable (pollutant) *X* at time *t*, and *L*—the lag distance operator, which superscript shows how many terms in the time series are taken back from the current time *t*. A time series represents an autoregressive model of order *p*, if it satisfies the *p*th degree difference equation

$$\begin{aligned} X_t &= \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \\ &= \left(\sum_{j=1}^{p} \phi_j L^j\right) X_t + \varepsilon_t, t = p+1,\ldots,n \end{aligned} \tag{3}$$

where $(\phi_1, \phi_2, \ldots, \phi_p)$ are constant parameters and ε_t is the error (white noise) at time *t*, assuming $\varepsilon_t \sim WN(0, \sigma^2)$. Note, that Eq. (3) has the form of regression equation for X_t, dependent on lagged values of itself (what is in fact an auto-regression).

Now consider the processes expressed by systematic fluctuations around some basic line. In this case, the current value X_t is a process represented by present and past values of white noise. A time series is a moving average model of order *q*, if it satisfies the *q*th degree difference equation

$$\begin{aligned} X_t &= \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \\ &= \left(1 - \sum_{j=1}^{q} \theta_j L^j\right) \varepsilon_t, t = q+1,\ldots,n \end{aligned} \tag{4}$$

where $(\theta_1, \theta_2, \ldots, \theta_q)$ are constant parameters.

In more complex cases, a trend may exists which is denoted by *d* and is usually determining as first, second or higher order differencing of X_t. For non-stationary time series with autoregressive and moving average processes, the general form of ARIMA (*p*, *d*, *q*) models has the form

$$\begin{aligned} &\left(1 - \sum_{j=1}^{p} \phi_j L^j\right)(1 - L)^d X_t = \left(1 - \sum_{j=1}^{q} \theta_j L^j\right) \varepsilon_t + c, \\ &t = \max(p+1, q+1), \ldots, n \end{aligned} \tag{5}$$

where *c* is a constant. Equation (5) includes the following parameters: $p, d, q, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$ and $\varepsilon_t \sim WN(0, \sigma^2)$.

If $\nabla = 1 - L$ is the differencing operator, we can write the model (5) as

$$\phi_p(L) \nabla^d X_t = \theta_q(L) \varepsilon_t + c. \tag{6}$$

Often in time series, a seasonal or periodical pattern also exists and is repeated every *s* observations. For monthly observations, $s = 12$ (12 in 1 year), for hourly observations $s = 24$ (24 in 1 day). In order to represent seasonality, ARIMA processes have been generalized to SARIMA models. The latter are formulated in the same way as (5) by including a term $(P, D, Q)_s$ with the same meaning as in ARIMA models. The general type SARIMA is then noted as ARIMA (*p*, *d*, *q*) $(P, D, Q)_s$, where *P* is the number of seasonal autoregressive (SAR) terms, *D* is the order of seasonal differencing and *Q* is the number of seasonal moving average (SMA) terms, respectively. In the seasonal part of the model, these three parameters operate across multiples of lag *s* (the number of periods in a season). In all cases, if a trend exists, the model does not include a constant term *c* (Tabachnik and Fidell 2005).

The general form of multiplicative seasonal ARIMA (*p*, *d*, *q*) $(P, D, Q)_s$ model is (Box and Jenkins 1976; Pankratz 1983):

$$\begin{aligned} &\phi_p(L) \Phi_P(L^s) \nabla^d \nabla_s^D X_t = \theta_q(L) \Theta_Q(L^s) \varepsilon_t + c, \\ &\quad \varepsilon_t \sim WN(0, \sigma^2), \end{aligned} \tag{7}$$

where ϕ, Φ are non-seasonal and seasonal autoregressive parameters (see also (6)), ∇ are differencing operators and θ, Θ are non-seasonal and seasonal moving average parameters. We have to note, that these parameters must lie within certain limits.

For example, the (1, 1, 1)(0, 1, 1)$_{24}$ model is expressed as $(1 - \phi_1 L) \nabla \nabla_{24} X_t = (1 - \theta_1 L)(1 - \Theta_{24} L^{24}) \varepsilon_t$ or in difference-equation form

$$\begin{aligned} (1 - \phi_1 L)(1 - L)\left(1 - L^{24}\right) X_t = {} & X_t - (1 + \phi_1) X_{t-1} \\ & + \phi_1 X_{t-2} - X_{t-24} \\ & + (1 + \phi_1) X_{t-25} - \phi_1 X_{t-26} \\ = {} & \varepsilon_t - \theta_1 \varepsilon_{t-1} - \Theta_{24} \varepsilon_{t-24} \\ & + \theta_1 \Theta_{24} \varepsilon_{t-25} \end{aligned} \tag{8}$$

Model (8) takes into account the dependence of a given value X_t on the value of the preceding two terms and the three past terms with a period of 24 h, as well as the moving average terms at the right hand side of the equation.

By neglecting the unknown value of ε_t, the model forecast $\hat{X}_t$ is

$$\hat{X}_t = (1+\phi_1)X_{t-1} - \phi_1 X_{t-2} + X_{t-24} - (1+\phi_1)X_{t-25} + \phi_1 X_{t-26} - \theta_1\varepsilon_{t-1} - \Theta_{24}\varepsilon_{t-24} + \theta_1\Theta_{24}\varepsilon_{t-25}. \tag{9}$$

Practical applications for fitting to actual data and forecasting depend on the facts that the univariate stochastic models yield forecasting that depend appreciably only on recent values of the series and forecasts are insensitive to small changes in parameter values with respect to the estimation errors (Box and Jenkins 1976).

5 Building models by the SARIMA methods

To build a SARIMA model, the Box–Jenkins empirical procedure requires the following steps: (1) Preliminary analysis: processing of data to satisfy the conditions of a Gaussian, stationary and invertible stochastic process; (2) Identification of the model by specifying the orders p, d, q, P, D, Q; (3) Estimation of model parameters; (4) Diagnosis of the appropriate goodness of fit measures, testing the parameters and residuals; (5) Application of the model: forecasting on holdout data sample and future cases, comparison and interpretation of the results.

This procedure is iterative in order to determine the most adequate model of the data on separate pollutants. When two or more models have almost equal fitting qualities the principle of parsimony is applied (Box and Jenkins 1976).

5.1 Preliminary analysis of the series for transformed data

This analysis was partially performed in Sect. 2. Usually, in order to determine whether the transformed variable represents a stationary process or not, a unit root test could be performed. We applied the augmented Dickey–Fuller test (ADF) (Said and Dickey 1984) to each of the transformed pollutant data samples, using the capabilities of EViews 7 software (EViews 7 for Windows 2013). This statistic is appropriate for large and more complicated sets of time series models and is applicable even for models of unknown order. The results of the unit root test at level 0.05 show, that all transformed variables are stationary ($d = 0$) with the exception of trO_3 ($d = 1$).

5.2 Identification of the model parameters for transformed data

5.2.1 Identification by ACF and PACF

A common tool for initial identification of the time series is the examination of their corresponding empirical autocorrelation functions (ACF) and partial autocorrelation functions (PACF). The patterns of ACF and PACF plots are used to find the appropriate model, describing its main behavior, including the presence of stationary process, trends, order of auto-regression and moving average processes, etc. (Tabachnik and Fidell 2005; Pankratz 1983). The ACF functions of the six transformed variables showed wave behavior, which indicated the presence of periodicity. In our case, this will lead to the availability of "seasonal" components in all time series, with respect to 24 h. By observing PACF, the presence of spikes outside the confidence limits was considered as an indication of approximate values of p and q, accordingly to the number of autoregressive (AR) and moving average (MA) terms in the model. It was found that $1 \le p \le 6$ and $1 \le q \le 9$ can be used as intervals for initial termination of these parameters for all variables. Despite the rough identification of the ARIMA in the case of more complex SARIMA models, the exact values of model parameters (p, q, P, D, Q) cannot be easily identified by using ACF and PACF functions and more precise techniques of identification are needed (Tabachnik and Fidell 2005).

5.2.2 Selection of model parameters by BIC information criterion

Often the number of parameters for an "optimal model" can be justified by using the more objective information criterions such as the AIC, BIC or others (Schwarz 1978; Burnham and Anderson 2002). The normalized Bayersian information criterion (BIC) is defined by

$$BIC = -2\frac{\ln(l_{\max})}{n} + \frac{k\ln(n)}{n} \tag{10}$$

where $l_{\max}$ is the maximum likelihood calculated for the model and k is the number of parameters of the model.

Among others, the preferred model is the one with the minimum BIC value. Note that this criterion is derived from various assumptions, including normal (Gaussian) or near-normal distribution of data (Liddle 2008), which in our case is fulfilled for transformed data.

For the variety of candidate models, the corresponding values of normalized BIC are shown in Table 5.

5.3 Estimation and diagnosis of the models for transformed variables

The extraction and estimation of model parameters is achieved using the ARIMA routines of the SPSS package. In addition to the BIC criterion, the root mean square error (RMSE) and the mean absolute percentage error (MAPE) have been used as more important model fit statistics. Note that RMSE is a good measure of accuracy used to compare forecasting errors of different models for a given variable and not between variables. The MASE is not taken into account in our investigation. Table 5 shows some of the examined candidate models and the results of model statistics, also including commonly used fit measures such as R^2 stationary, R^2, and the significance of the models.

The final selected models of the six air pollutants are marked by (*) in Table 5, in the first row next to the corresponding transformed variable. The selection of the models was performed under the following combined criteria: (1) Minimum BIC; (2) Minimum RMSE; (3) Maximum R^2; (4) Minimum significance Sig.; (5) Minimum MAPE. Some of the other candidate models sorted by this criteria are listed below the selected optimal models.

The distributions of residuals within 5 % confidence intervals and the normality of residuals for all candidate models were checked before applying the proposed combined criteria procedure.

5.4 Forecasting

The strength of the ARIMA and SARIMA models lies in the good results achieved when forecasting future events (Chatfield 2000, Pankratz 1983). In our case, we have presented the application of the obtained SARIMA models in a short-term ahead forecasting (within 72 h) as of 00:00 on 1st September 2012 up to 24:00 of 3rd September 2012. This time period follows directly the date used in the models, i.e. actual additional data are used which have not been included in the construction of the models therefore these can be compared to the predictions made using the models. The period of 3 days is a standard for this type of forecasting and is usually between 2 and 5 days in duration. Over longer forecasting periods, in our case, the accuracy of the models is not satisfactory.

In Figs. 1, 2, 3, 4, 5, and 6 the model fitted values for each of the six pollutants over the period of 1 year have been highlighted in blue. As it is shown, there is very good correspondence with the pattern of changes.

Separate plots of the last 72 h observed data and 72 h forecast results (72 h outside the investigated data) compared with the measured holdout real data are also given in Figs. 7, 8, 9, 10, 11, and 12 for all pollutants. More details are discussed in the next section.

6 Analysis and interpretation of the results

6.1 Discussion of factor analysis results

First, we interpret the results from the factor analysis. The correlation matrix (see Table 3) indicates the presence of highly bicorrelate relationships between the concentrations of NO, NO_2, NO_x and PM10. This means that these pollutants have almost the same overall behavior—they increase or decrease simultaneously over time, which may make their management easier. Correlations between different air pollutants are found relatively rarely in literature. An example is (Kumar and Joseph 2006), where high correlation has been found between PM10, PM2.5 and NO_2, and (Ko et al. 2007), investigating correlations between SO_2, NO_2, PM10, O_3 and PM2.5.

In this study, the explicit grouping of the six investigated pollutants in three factors can be explained by the presence of the same defining common causes of pollution. As expected, factor $F1$ groups together the pollutants NO_2, NO, PM10 and NO_x. They are believed to be caused by the use of solid fuels (including coal combustion) by households and the thermal power stations located within the town and have higher levels in winter, confirmed in previous periods by the EAA agency (EAA 2013). The contribution of NO and NO_x dominates in the factor. The two other pollutants (PM10, NO_2), which show almost the same behavior in their time series, have been observed to peak in winter. However, by examining the separate plot of PM10 in Fig. 4, it is evident that this pollutant systematically surpasses the official hour limit norm of 50 μg/m^3 and the average threshold limit per year (see Table 1). We also have to add that small Bulgarian towns do not have significant road traffic and the main sources of pollution are households due to the lack of centralized heating. This is the case with small towns elsewhere in Europe, too. For example, in south western Poland it has been found that local combustion sources contributed up to 80 % to PM10 mass concentration in winter (Zwozdziak et al. 2012).

Factor $F2$ includes only the pollutant ozone (O_3), which correlates negatively with nitrogen oxides but its presence is localized and its levels do not exceed the admissible values for this region and the country, except on some very hot summer afternoons. However, this pollutant shows a positive trend and has to be monitored with caution. Sulfur dioxide SO_2 is separated in $F3$ but its values for the town of Blagoevgrad do not exceed the national and European prescribed limits and standards (EEA 2013; Reports and Bulletins 2012; Directive 2008; Air Quality Standards 2013). Its explicit separation from the other pollutants is easily attributed to its main source—the moderate level of road traffic in the town (EEA 2013; Reports and Bulletins 2012).

Another crucial advantage of the application of factor analysis is finding the actual ratio between the main groups

Table 5 SARIMA models of air pollutants (transformed data) for city of Blagoevgrad with model fit statistics

Transformed variable	SARIMA model	Model fit statistics					
		Stationary *R*-squared	*R*-squared	RMSE	MAPE	Normalized BIC	Sig.
trNO_2	*$(2,0,1)(2,0,1)_{24}$	0.841	0.841	0.290	8.461	−2.467	0.003
	$(3,0,3)(2,0,2)_{24}$	0.841	0.841	0.290	8.458	−2.461	0.000
	$(3,0,2)(2,0,1)_{24}$	0.839	0.839	0.292	8.519	−2.452	0.000
	$(2,0,1)(1,0,1)_{24}$	0.839	0.839	0.293	8.525	−2.451	0.000
	$(4,0,3)(2,0,1)_{24}$	0.839	0.839	0.292	8.533	−2.449	0.000
trNO	*$(1,0,7)(2,0,1)_{24}$	0.783	0.783	0.451	225.84	−1.579	0.003
	$(2,0,7)(2,0,1)_{24}$	0.783	0.783	0.451	225.72	−1.577	0.001
	$(1,0,6)(2,0,1)_{24}$	0.783	0.783	0.452	233.43	−1.577	0.000
	$(1,0,6)(1,0,1)_{24}$	0.782	0.782	0.452	232.46	−1.576	0.000
	$(0,0,8)(2,0,1)_{24}$	0.779	0.779	0.456	232.26	−1.559	0.000
trO_3	*$(2,1,1)(1,1,1)_{24}$	0.437	0.941	3.740	13.813	2.643	0.000
	$(3,1,1)(1,1,1)_{24}$	0.437	0.904	3.740	13.810	2.644	0.000
	$(3,1,2)(1,1,1)_{24}$	0.437	0.941	3.739	13.778	2.645	0.000
	$(2,1,2)(2,0,1)_{24}$	0.394	0.940	3.741	13.833	2.645	0.003
	$(2,1,2)(1,0,1)_{24}$	0.394	0.940	3.741	13.833	2.645	0.003
	$(2,1,1)(1,0,1)_{24}$	0.360	0.937	3.845	13.597	2.699	0.000
trPM10	*$(4,0,4)(1,0,1)_{24}$	0.888	0.888	0.142	4.170	−3.897	0.000
	$(3,0,4)(1,0,1)_{24}$	0.888	0.888	0.142	4.163	−3.896	0.000
	$(4,0,4)(3,0,1)_{24}$	0.888	0.888	0.142	4.169	−3.896	0.000
	$(4,0,4)(1,0,3)_{24}$	0.888	0.888	0.142	4.168	−3.896	0.000
	$(3,0,3)(1,0,1)_{24}$	0.887	0.887	0.142	4.167	−3.893	0.000
	$(3,0,2)(1,0,1)_{24}$	0.887	0.887	0.142	4.17	−3.893	0.000
trSO_2	*$(2,0,2)(2,0,1)_{24}$	0.887	0.887	0.369	20.492	−1.988	0.000
	$(3,0,2)(2,0,1)_{24}$	0.887	0.887	0.368	20.42	−1.987	0.000
	$(3,0,1)(3,0,1)_{24}$	0.886	0.886	0.369	20.495	−1.986	0.000
	$(4,0,1)(2,0,2)_{24}$	0.886	0.886	0.37	22.017	−1.978	0.000
	$(2,0,1)(2,0,1)_{24}$	0.885	0.885	0.371	22.177	−1.978	0.000
	$(2,0,3)(2,0,1)_{24}$	0.878	0.878	0.382	22.84	−1.916	0.000
trNO_x	*$(3,0,2)(2,0,1)_{24}$	0.841	0.841	0.191	8.595	−3.302	0.009
	$(3,0,2)(1,0,1)_{24}$	0.842	0.842	0.191	8.611	−3.302	0.004
	$(3,0,1)(2,0,1)_{24}$	0.842	0.842	0.191	8.599	−3.301	0.020
	$(4,0,3)(2,0,3)_{24}$	0.842	0.842	0.191	8.590	−3.298	0.001
	$(3,0,2)(0,0,1)_{24}$	0.826	0.826	0.200	9.089	−3.208	0.000
	$(2,0,2)(2,0,1)_{24}$	0.811	0.811	0.209	9.331	−3.121	0.000

Selected optimal models are denoted by *

of pollutants and defining the ones which influence air quality the most, and which should be subjected to the most stringent control by the responsible authorities. In our case, the most significant factor for air pollution in Blagoevgrad is the complex one—*F*1 with an influence of 45 %, following by 27 % for ozone and 19 % for SO_2.

6.2 Discussion of time series analysis results

The second approach in this study has the aim of modeling the investigated air pollutants with respect to time by using SARIMA method.

The selected SARIMA models between the set of all considered candidate models have relatively simple form and demonstrate sufficiently good model fit statistics (see Table 5). Trend and seasonal trend are available only for O_3. All models show high values of *R* squared, which means that the models describe significant parts of the data. All these models are significant and residuals are normally distributed and vanish to zero. This way, the obtained SARIMA models for the transformed data of the pollutants can be accepted as the best models.

The results for short-term forecasting within 72 h are given in the right hand sides of the vertical lines in Figs. 7,

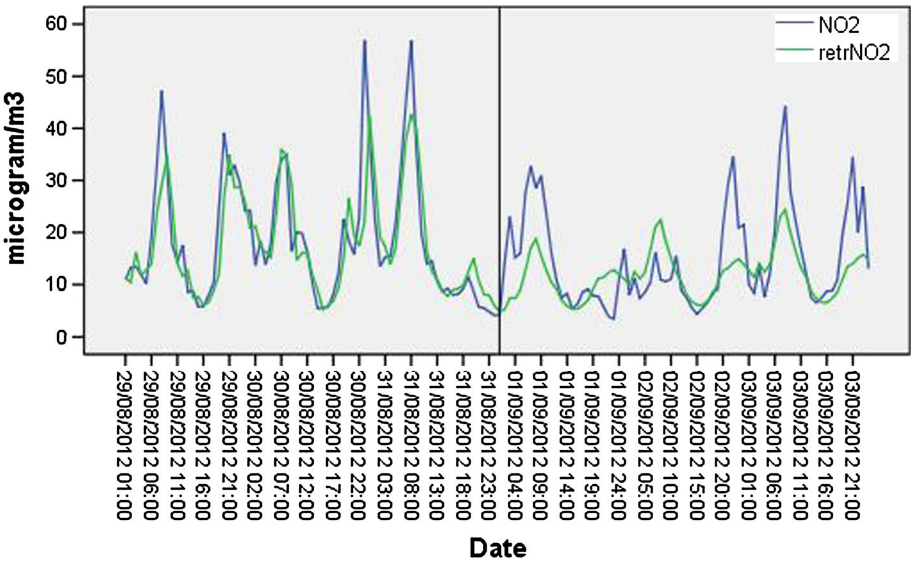


Fig. 7 Comparison of the observed data in the last 72 h (*at the left hand side of the vertical line*) and a forecasting for NO_2 using holdout real data for 72 h (*at the right hand side of the vertical line*) with the retransformed SARIMA model *(2,0,1)(2,0,1)$_{24}$ (retrNO_2)

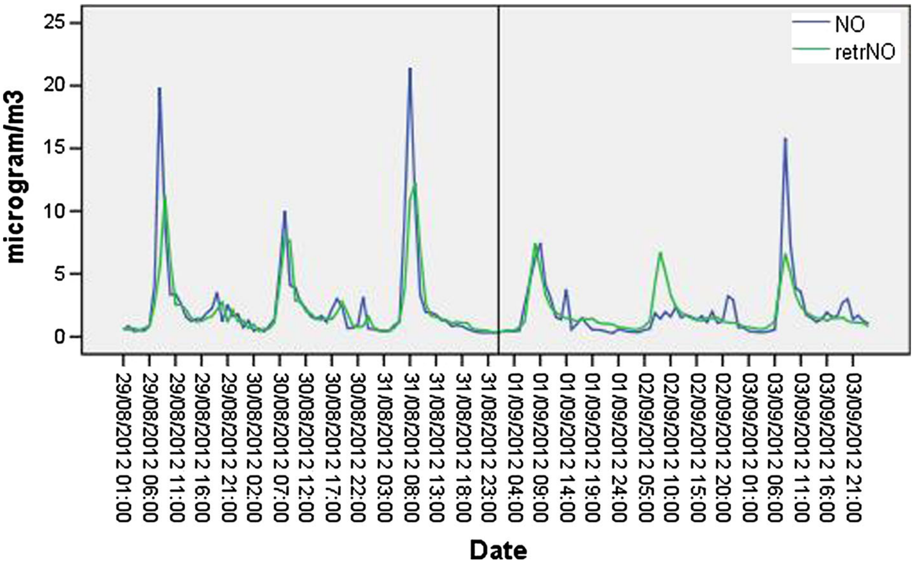


Fig. 8 Comparison of the observed data in the last 72 h (*at the left hand side of the vertical line*) and a forecasting for NO using holdout real data for 72 h (*at the right hand side of the vertical line*) with the retransformed SARIMA model *(1,0,7)(2,0,1)$_{24}$ (retrNO)

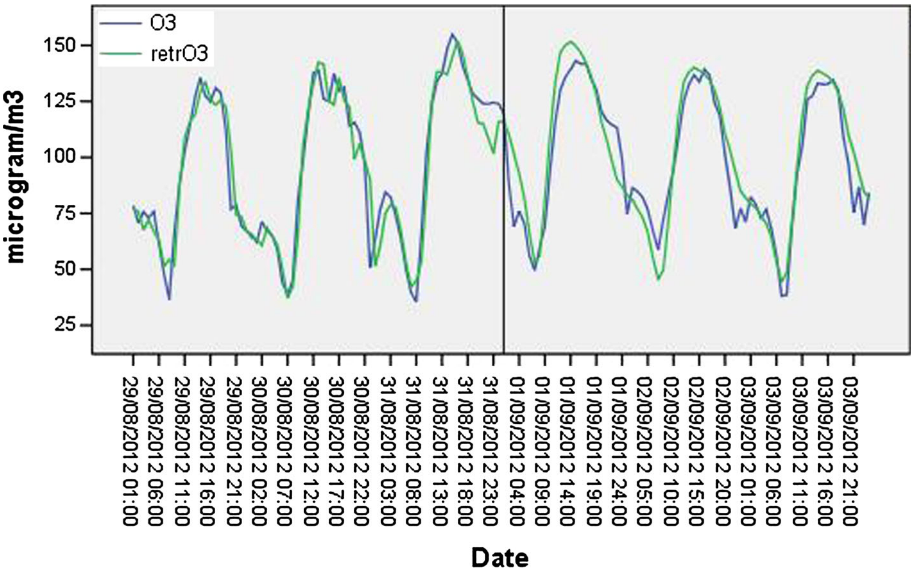


Fig. 9 Comparison of the observed data in the last 72 h (*at the left hand side of the vertical line*) and a forecasting for O_3 using holdout real data for 72 h (*at the right hand side of the vertical line*) with the retransformed SARIMA model *(2,1,1)(1,1,1)$_{24}$ (retrO_3)

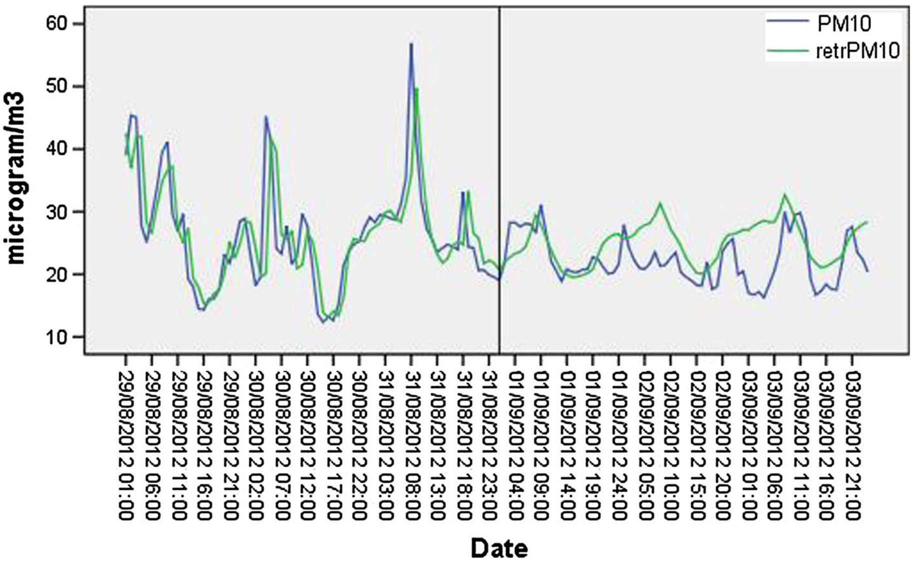


Fig. 10 Comparison of the observed data in the last 72 h (*at the left hand side of the vertical line*) and a forecasting for PM10 using holdout real data for 72 h (*at the right hand side of the vertical line*) with the retransformed SARIMA model $*(4,0,4)(1,0,1)_{24}$ (retrPM10)

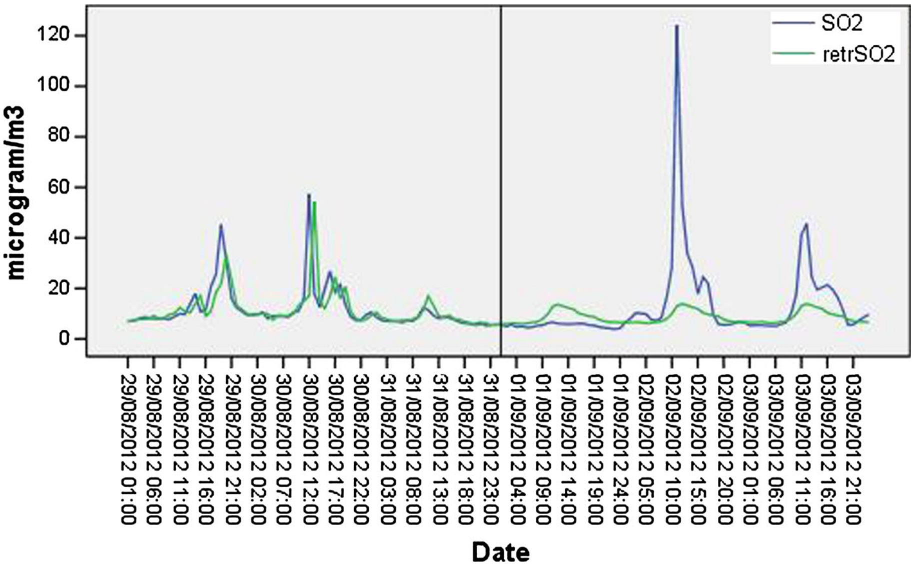


Fig. 11 Comparison of the observed data in the last 72 h (*at the left hand side of the vertical line*) and a forecasting for SO_2 using holdout real data for 72 h (*at the right hand side of the vertical line*) with the retransformed SARIMA model $*(2,0,2)(2,0,1)_{24}$ (retrSO_2)

8, 9, 10, 11, and 12 and show very good predictive properties, useful for practical issues, like preventing future over-pollution, caution alarms or even for more extended periods of time. Specifically, Figs. 7, 8, 10 and 12 show good predictions in cases with drops and peaks. Figure 9 shows very good correspondence for the level of ozone pollution, which is attributed to the high coefficient of determination R^2 and the model having identified a trend. The overall comparison demonstrates that the weakest prediction quality is that for SO_2 given in Fig. 11, where the sharp peaks have not been predicted sufficiently well. The left side of Fig. 11 shows the excellent performance of the model for fitting the observed pollutant concentrations during the last 72 h of the considered time period. What is more, for the last day, these have almost the same low values. The obtained forecasts for the next day are also excellent when compared to actual observations other than the ones used in the model, retaining the same behavior pattern. According to the model formula $*(2,0,2)(2,0,1)_{24}$, the calculated predicted values are dependent on the low values during the previous 27 h (see (9) for comparison). For this reason, the model has poor performance characteristics when compared against the observed peak of SO_2 pollution. However, the maximum peak value is within the threshold limit of 125 µg/m3 per day (Table 1). This indicates a certain flaw in the applied methodology which uses parametric models for time series (including for ARIMA and SARIMA methods) of data with sharp peaks and drops.

Finally, we have to add that a large number of additional investigations were carried out using the two considered approaches (factor analysis and SARIMA methods) applied

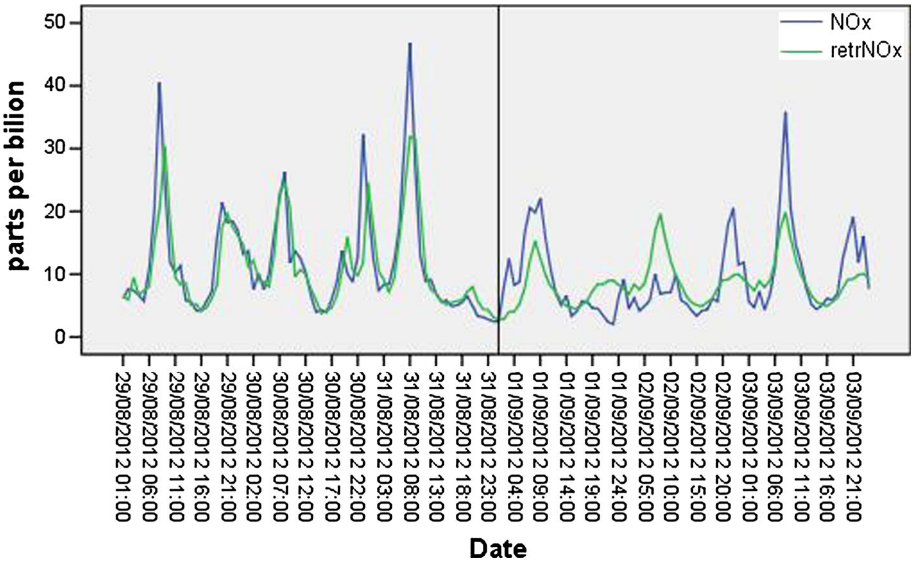


Fig. 12 Comparison of the observed data in the last 72 h (*at the left hand side of the vertical line*) and a forecasting for NO$_x$ using holdout real data for 72 h (*at the right hand side of the vertical line*) with the retransformed SARIMA model $*(3,0,2)(2,0,1)_{24}$ (retrNO$_x$)

to the data for Blagoevgrad in order to study other time periods with duration of 1 year beginning at different initial dates. The predictions for these were similar in terms of accuracy, and the results presented here are only one variant. Based on this, it can be concluded that the application of both approaches yields very good stable results, which do not depend significantly on the period being studied.

7 Conclusion

This study presents a statistical investigation of six pollutants of the ambient air quality in the town of Blagoevgrad, a small typical town of Bulgaria. Two statistical approaches are applied to model and predict the observed actual data over a period of 1 year, based on hourly measurements.

By applying factor analysis, strong correlations were found between the various air pollutants, based on which six air pollutants were grouped down into three factors and the degree of influence of each factor in the overall pollution pattern was determined. The three factors identified mixed effects of pollution. This has been interpreted as the presence of common sources for the pollutants in the obtained groups.

The main part of the paper is related to the derivation and application of univariate Box–Jenkins stochastic SARIMA models for any of the six pollutants. A positive first degree trend was established for ozone pollution. In particular, the results indicated that PM10 concentrations tend to be higher in winter due to the residential wood burning as a major pollution source. The values of PM10 exceed the official national and European norms, so that the status of this ecological indicator is highly troubling. The models were implemented for short-term forecasting for a future period of 72 h and the results demonstrate sufficiently good performance compared with the real data. The best models were obtained and selected on the basis of a BIC information criterion and other commonly used goodness of fit criteria.

A significant moment in the modeling of time series is the use of the Yeo–Johnson power transformation for variance stabilizing of the data which led to the development of relatively non-complex univariate stochastic models with very good statistical indices. Furthermore, recording on an hourly basis provided very good results when fitting the observed concentrations of the different air pollutants and short-term forecasts, in particular those for ozone and particulate matter PM10.

Overall, it was shown that factor analysis and the SARIMA approach are very appropriate tools for examining air pollution levels in small urban areas in order to provide assistance in everyday control and forecasting of the air quality. The future goal of the investigation will be to build non-parametric models and examine their ability to improve forecasting.

The town of Blagoevgrad gives example of a typical small and medium urban region in basin valleys of Bulgaria. The specific geographic conditions in complex with the everyday urban activities result in unsatisfactory air quality, according to last year monitoring data. Hence, these regions demand more efficient control and forecasting procedure for minimize and avoid the ascertained exceeding of PM10 and partially other pollutants.

This paper analyzes the current status of air quality during 1 year period and demonstrates the relevant tools for effective statistical analysis and forecasting the levels of the main air pollutants in such type of urban regions.

Acknowledgments This paper is published in cooperation with project of the Bulgarian Ministry of Education, Youth and Science, BG051PO001/3.3-05-0001 "Science and business", financed under Operational program "Human Resources Development" by the European Social Fund and is also partially supported by the Scientific Research Department of Plovdiv University Paisii Hilendarski under Grant NI13 FMI-002.

References

Air Quality Standards (2013) European Commission. Environment. http://ec.europa.eu/environment/air/quality/standards.htm. Accessed 26 April 2013

Blagoevgrad (2013) Wikipedia. http://en.wikipedia.org/wiki/Blagoevgrad. Accessed 26 April 2013

Blifford IH Jr, Meeker GO (1967) A factor analysis model of large scale pollution. Atmos Environ 1:147–157. doi:10.1016/0004-6981(67)90042-X

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc Ser B 26:211–252

Box GEP, Jenkins GM (1976) Time series analysis, forecasting and control, revized edn. Holden Day, San Francisco

Brunelli U, Piazza V, Pignato L, Sorbello F, Vitabile S (2007) Two-days ahead prediction of daily maximum concentrations of SO_2, O_3, PM10, NO_2, CO in the urban area of Palermo, Italy. Atmos Environ 41(11):2967–2995. doi:10.1016/j.atmosenv.2006.12.013

Burnham KP, Anderson DR (2002) Model selection and inference: a practical information-theoretic approach, 2nd edn. Springer, New York

Chatfield C (1996) The analysis of time series: an introduction. Chapman & Hall/CRC, Boca Raton

Chatfield C (2000) Time-series forecasting. Chapman & Hall CRC, Boca Raton

Comparison of Statistical Packages. http://en.wikipedia.org/wiki/Comparison_of_statistical_packages. Accessed 15 July 2013

Díaz-Robles LA, Ortega JC, Fu JS, Reed GD, Chow JC, Watson JG, Moncada-Herrera JA (2008) A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. Atmos Environ 42(35): 8331–8340. doi:10.1016/j.atmosenv.2008.07.020

Dimov I, Georgieva R, Ivanovska S, Ostromsky T, Zlatev Z (2010) Studying the sensitivity of pollutants' concentrations caused by variations of chemical rates. J Comput Appl Math 235:391–402. doi:10.1016/j.cam.2010.05.041

Dimov I, Georgieva R, Ostromsky T, Zlatev Z (2013) Advanced algorithms for multidimensional sensitivity studies of large-scale air pollution models based on Sobol sequences. Comput Math Appl 65(3):338–351. doi:10.1016/j.camwa.2012.07.005

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. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:152:0001:0044:EN:PDF. Accessed 26 April 2013

Dueñas C, Fernández MC, Cañete S, Carretero J, Liger E (2005) Stochastic model to forecast ground-level ozone concentration at urban and rural areas. Chemosphere 61(10):1379–1389. doi:10.1016/j.chemosphere.2005.04.079

EEA Daily Bulletin for air quality in the country (2013) EEA—Executive environment agency, National system for realtime air quality control in Bulgaria. http://pdbase.government.bg/airq/bulletin-en.jsp. Accessed 26 April 2013

EEA Reports, Bulletins (2012) EEA—Executive environment agency. http://eea.government.bg/en/output/index.html. Accessed 26 April 2013

EViews 7 for Windows (2013) http://www.eviews.com. Accessed 26 April 2013

Gardner MW, Dorling SR (1999) Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. Atmos Environ 33(5):709–719. doi:10.1016/S1352-2310(98)00230-1

Henry RC, Hidy GM (1979) Multivariate analysis of particulate sulfate and other air quality variables by principal components—part I: annual data from Los Angeles and New York. Atmos Environ 13(11):1581–1596. doi:10.1016/0004-6981(79)90068-4

Huang J, Ho M, Du P (2011) Assessment of temporal and spatial variation of coastal water quality and source identification along Macau peninsula. Stoch Environ Res Risk A 25(3):353–361. doi:10.1007/s00477-010-0373-4

ISO/IEC 17025 (2013) Wikipedia. http://en.wikipedia.org/wiki/ISO/IEC_17025. Accessed 26 April 2013

Ivanov A, Voynikova D, Gocheva-Ilieva S, Boyadzhiev D (2012) Parametric time series analysis of daily air pollutants of city of Shumen, Bulgaria. AIP Conf Proc 1487:386–396. American Institute of Physics, Melville, NY. doi:10.1063/1.4758982

Jacobson MZ (2005) Fundamentals of atmospheric modeling, 2nd edn. Cambridge Univ. Press, Cambridge

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

Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York

Kaplunovsky AS (2005) Factor analysis in environmental studies. HAIT J Sci Eng B 2(1–2):54–94

Kim SE (2010) Tree-based threshold modeling for short-term forecast of daily maximum ozone level. Stoch Environ Res Risk A 24(1):19–28. doi:10.1007/s00477-008-0295-6

Kim SE, Kumar A (2005) Accounting seasonal nonstationarity in time series models for short-term ozone level forecast. Stoch Environ Res Risk A 19:241–248. doi:10.1007/s00477-004-0228-y

Kim J, Mueller CW (1986) Factor analysis: statistical methods and practical issue. Sage, Beverly Hills

Ko FWS, Tam W, Wong TW, Chan DPS, Tung AH, Lai CKW, Hui DSC (2007) Temporal relationship between air pollutants and hospital admissions for chronic obstructive pulmonary disease in Hong Kong. Thorax 62:780–785. doi:10.1136/thx.2006.076166

Kumar U, Jain VK (2010) ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO). Stoch Environ Res Risk A 24:751–760. doi:10.1007/s00477-009-0361-8

Kumar R, Joseph AE (2006) Air pollution concentrations of PM2.5, PM10 and NO_2 at Ambient and Kerbsite and their correlation in Metro City—Mumbai. Environ Monit Asses 119(1–3):191–199. doi:10.1007/s10661-005-9022-7

Lengyel A, Heberger K, Paksy L, Banhidi O, Rajko R (2004) Prediction of ozone concentration in ambient air using multivariate methods. Chemosphere 57(8):889–896. doi:10.1016/j.chemosphere.2004.07.043

Liddle AR (2008) Information criteria for astrophysical model selection. Cornell University Library, arHiv.org. http://arxiv.org/PS_cache/astro-ph/pdf/0701/0701113v2.pdf. Accessed 26 April 2013

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. doi:10.1016/j.atmosenv.2009.01.055

McBerthouex P, Brown LC (2002) Statistics for environmental engineers. Lewis Publishers, Boca Raton

Milionis AE, Davies TD (1994) Regression and stochastic models for air pollution. I. Review, comments and suggestions. Atmos Environ 28:2801–2810. doi:10.1016/1352-2310(94)90083-3

National System for Environmental Monitoring, Bulgaria (2013) http://eea.government.bg/en/nsmos/index.html. Accessed 26 April 2013

Pankratz A (1983) Forecasting with univariate Box–Jenkins models: concepts and cases. Wiley, New York

Petelin D, Grancharova A, Kocijana J (2013) Evolving Gaussian process models for prediction of ozone concentration in the air. Simul Model Pract Th (EUROSIM 2010), 33:68–80. doi:http://dx.doi.org/10.1016/j.simpat.2012.04.005

Polydoras GN, Anagnostopoulos JS, Bergeles GCh (1998) Air quality predictions: dispersion model vs Box-Jenkins stochastic models. An implementation and comparison for Athens, Greece. Appl Therm Eng 18(11):1037–1048. doi:10.1016/S1359-4311(98)00016-7

Program for reducing the harmful air emission of PM10 in the territory of Blagoevgrad (2011) Municipality of Blagoevgrad. http://eea.government.bg/bg/nsmos/air/roukav/obshtini2/KAV-Blagoevgrad.doc (in Bulgarian)

Richman MB (1986) Rotation of principal components. J Climatol 6:293–335. doi:10.1002/joc.3370060305

Said ES, Dickey DA (1984) Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika 71(3): 599–607. doi:10.1093/biomet/71.3.599

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464. doi:10.1214/aos/1176344136

Sharma P, Chandra A, Kaushik SC (2009) Forecasts using Box–Jenkins models for the ambient air quality data of Delhi City. Environ Monit Assess 157(1–4):105–112. doi:10.1007/s10661-008-0520-2

Slini Th, Karatzas K, Moussiopoulos N (2002) Statistical analysis of environmental data as the basis of forecasting: an air quality application. Sci Total Environ 288:227–237. doi:10.1016/S0048-9697(01)00991-3

SPSS IBM Statistics (2013) http://www-01.ibm.com/software/analytics/spss/. Accessed 26 April 2013

Tabachnik BG, Fidell LS (2005) Using multivariate statistics, 5th edn. Pearson Int. Edition, Boston

Yeo IK, Johnson RA (2000) A new family of power transformations to improve normality or symmetry. Biometrika 87(4):954–959. doi:10.1093/biomet/87.4.954

Zwozdziak A, Samek L, Sowka I, Furman L, Skrętowicz M (2012) Aerosol pollution from small combustors in a village. Sci World J 2012: Article ID 956401. doi:10.1100/2012/956401