ECOLOGIA BALKANICA

2016, Vol. 8, Issue 2 December 2016 pp. 33-41

Stochastic Modeling of Problematic Air Pollution with Particulate Matter in the City of Pernik, Bulgaria

Maya P. Stoimenova[*]

University of Plovdiv „Paisii Hilendarski“, Faculty of Mathematics and Informatics,
Department of Applied Mathematics and Modeling
24 Tzar Assen Str., 4000 Plovdiv, BULGARIA
*Corresponding author: mstoimenova@uni-plovdiv.bg

Abstract. Air quality in urban areas is an important prerequisite for a healthy environment. This paper focuses on the study of the problematic pollutant PM10 in the air over the city of Pernik in order to prevent the worsening of air pollution and to meet the requirements of the applicable regulations and directives, as well as to improve public awareness with regard to health and environmental issues. In this paper, stochastic mathematical models are developed using average 24-hour concentrations of PM10 in atmospheric air over the city for the period from 1 January 2010 until 31 December 2014. The measured values systematically exceed European Union regulations with require that mean daily concentrations should be below 50 μg/m^3. Univariate time-dependent models are derived in the form of time series. The constructed models describe the examined data adequately and also make it possible to forecast future pollution within a timeframe of several days. The selected type of modelling facilitates the decision making needed in the efforts to decrease the pollution levels in future.

Key words: Particulate matter PM10, air pollution, stochastic model.

Introduction

It is well-known that PM10 (particles with an aerodynamic diameter between 2.5 and < 10 μm) can be harmful to human health, which has been shown by epidemiological analysis and toxicological investigations (WHO, 2013). These particles are produced mainly by industrial factories, motor vehicles, and households burning solid fuels, and they can cause asthma, cardiovascular disease, lung cancer, and in some cases, even premature death. The concentrations of PM10 are regulated and monitored in accordance with the provisions of the European Union (EU) as set out in Directive 2008/50/EC (2008) and the air quality standards by the European Commission (2015), pursuant to which the maximum permissible values for PM10 in the air include mean annual values of 40 μg/m^3 and a 24-hour mean value of 50 μg/m^3, which must not be exceeded more than 35 times within one calendar year. Due to the adverse effects on human health caused by PM10 and in order to comply with EU regulations, it is necessary to monitor whether these provisions are met and to undertake measures to eliminate and prevent any violations. In Bulgaria, systematic air pollution with PM10 is problematic in many towns and cities with the worst affected being Pernik (the most polluted one in Europe in 2011), Plovdiv, and Sofia among others, with these often ranking as some of

Union of Scientists in Bulgaria – Plovdiv
University of Plovdiv Publishing House

the urban areas with the highest concentrations of this pollutant in Europe during the last few years (WHO, 2015).

Observation, control, and forecasting of air pollutant concentrations in Bulgaria are performed by the Executive Environment Agency, and the results of its work are published on the website (ExEA, 2016), along with regular air quality reports for all regions within the country (EEA, 2016).

In addition to the official means to study air pollution, publications also apply various mathematical approaches for modeling accumulated statistical data. A large group of these are stochastic modeling methods for time series, encompassing mostly the ARIMA, SARIMA, and transfer function methods, as well as others. They are based on the Box-Jenkins methodology (Box *et al.*, 1994), which in addition to environmental studies, also finds a wide range of applications in other fields such as economics, econometrics, demographics, etc. ARIMA models in environmental modeling literature have been used, for example in (Sharma *et al.*, 2009; Slini *et al.*, 2006). Hybrid ARIMAs are also applied in combination with the artificial neural network methods, multivariate regression, principle component method, and others, in order to investigate PM10 concentrations (see for example Vlachogianni *et al.* 2011; Vesely *et al.*, 2009; Zwozdziak *et al.*, 2012; Stadlober *et al.*, 2012).

There are very few publications related to the mathematical modeling of PM10 air pollution in urban areas within Bulgaria. We have to note the recent papers (Gocheva-Ilieva *et al.*, 2014; Ivanov *et al.*, 2015; Ivanov & Gocheva-Ilieva, 2013; Voynikova *et al.*, 2015), wherein various methods are applied, such as stochastic modeling, factor and regression analysis, intelligent machine learning methods, etc.

This paper examines the status of air pollution with PM10 in the city of Pernik over a period of 5 years based on average daily measurements. The objective is to construct suitable stochastic models describing actual data, to analyze the quality of the models, and to select an adequate one, as well as to apply the model for estimation and forecasting of future pollution within a timeframe of several days. The availability of data for a relatively long period allows for the application of a method for finding long-term trends, as well as making short-term forecasts for future pollution in order to warn local authorities and the public about any dangers resulting from the exceeded permissible limits for particulate matter concentration.

Data analysis was performed using the statistical software SPSS (IBM Corp., 2013).

Material and Methods

Study area. Pernik is among Bulgaria's cities with the worst air quality, due to the presence of harmful pollutants. The city is located in the south-west of Bulgaria, on the banks of the Struma River. The climate is temperate continental and the average altitude is 750 meters. Rainfall is markedly continental in nature which facilitates pollution and air self-cleaning processes. In 2015, the city's population was around 75,000 people, making it the second most populous city in western Bulgaria after the capital Sofia, and 11th in the country. Pernik is the largest city in Bulgaria where coal is mined. Pan-European transport corridors 4 and 8 pass near the city along the “Lyulin” and “Struma” Motorways, as well as European road E871 and the railway line connecting central Europe with Greece. Among the main sources of air pollution, in addition to vehicle traffic, are large factories such as "Stomana" AD steelworks, fossil-fuel power station "Republika", etc.

Data and initial statistical processing. Our analysis of the situation in Pernik is based on data measured at the "Shahtyor" station, located in close proximity to the city center. The station is equipped with analyzers for continuous measurement of the main and specific atmospheric pollutants such as nitrogen monoxide, nitrogen dioxide, carbon monoxide, carbon dioxide, benzene, PM2.5, and PM10. The data we used cover a period of 5 years – from 1 January 2010 until 31 December 2014 based on average daily observations.

ARIMA Method. ARIMA (Auto-Regressive Integrated Moving Average) models are noted as ARIMA (p,d,q) (BOX *et al.*, 1994). The autoregression element p represents the influence of the data at each moment t of p previous moments within the model. The integrated element d represents trends in the data, and the element q indicates the number of members used to construct the small fluctuations with the help of a moving average.

The main steps of ARIMA methods are:

- Identification – examining the data along with calculation and drawing a graph of auto-correlation functions (ACF) and partial auto-correlation functions (PACF). The smallest values of the parameters are sought. When the value is 0, the element is not needed in the respective model. The element d (trend) is examined first. The goal is to determine whether the process is stationary (d=0), and if it is not, to be transformed into such. The value of p is 0, if there is no connection between every two sequential observations.
- Constructing models and estimating its parameters.
- Diagnostics and selection of model – the residuals and the quality of approximation of the model are examined. The residuals are the differences between the values predicted by the model and the observed data. Theoretically, it is assumed that the residuals are random and normally distributed.
- Application of the predictive model, forecasts, analysis of dependencies, and study problem-solving capabilities.

Results

The results of the initial processing of the data are given in Table 1. As shown, there are no missing values in the measurements. Number of observations N=1826.

The maximum value of 348 μg/m³ for PM10 exceeds 7 times the average daily limit to 50 μg/m³. Such excesses are not isolated occurrences. Table 1 shows that the average value of PM10 of 63,180 μg/m³ for the 5 years exceeds the upper limit for the annual value of 40 μg/m³ as stipulated in European and Bulgarian legislation. Fig. 1 characterizes the behavior of the concentrations of the pollutant over time.

It is known that, the application of parametric models requires normal or close to normal distribution of time variables (WILKS, 2011). The distribution for our original data is shown in Fig. 2 a. In order to improve the distribution and to minimize the variability of the data before constructing the models, various transformations of the initial data are widely used in ecology (WILKS, 2011; BOX *et al.*, 1994).

The initial transformation of the data is performed using the formula (YEO & JOHNSON, 2000):

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

where x is the initial variable, *trx* is the transformed variable, and λ is an unknown parameter. For our data, the optimal transformation parameter λ is determined using the simple procedure through tests from the values [-2, -1.9, …, 2] and the Kolmogorov-Smirnov test of normality. The effect of the applied transformation is illustrated in Fig. 2 b) – the distribution is close to normal with the chosen optimal parameter λ=*-0.4.*

Table 1. Descriptive statistics of the initial data on PM10 (variable SPM0_1) concentrations in the city of Pernik.

Mean	Median	Std. Dev-iation	Var-iance	Skew-ness	Std. Error of Skewness	Kur-tosis	Std. Error of Kurtosis	Min.	Max.
63.180	47.700	50.273	2527.360	2.272	0.057	5.895	0.114	7	348

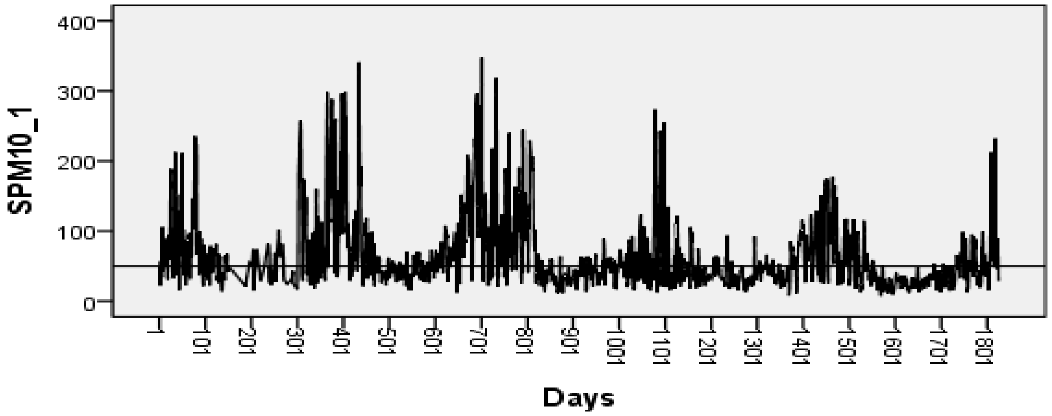
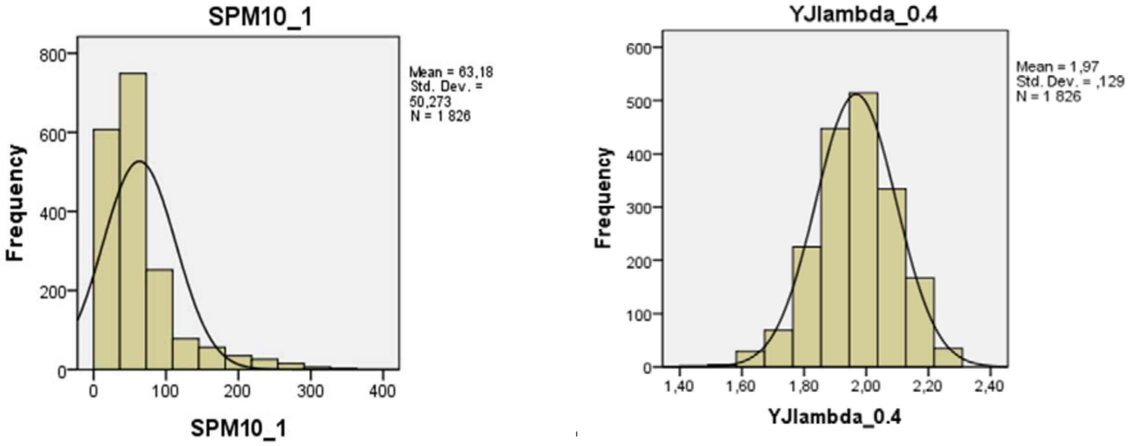


Fig. 1. Graph of the measured average daily data on PM10 concentrations in the city of Pernik. The horizontal line indicates the permissible limit of 50 µg/m³.



a) b)

Fig. 2 a, b. Distribution graphs of the initial data of PM10 for the city of Pernik: a) before the transformation; b) after the Yeo-Johnson transformation with *λ=-0.4*.

Construction of ARIMA models and analysis using transformed time series

As part of the analysis of time series, the ACF and PACF functions are used to identify periods and trends in the data. The distribution in the graph of ACF and PACF coincides with the theoretically perfect distribution of a more suitable model describing the main behavior, including the presence of a stationary process, linear or quadratic trends, levels of autoregression, and moving averages, etc.

For the construction of the parametric models, we apply the ARIMA method, looking for models in the form ARIMA(p,d,q) (Box *et al.*, 1994). In Fig. 3 a, ACF goes down slowly, and the PACF function in Fig. 3 b demonstrates three dramatic peaks with delay lags 1, 3, and 6, as well as characteristic distribution with delay in lag 7. This means the PM10 model may be found with three lags back in autoregression (AR) and moving averages (MA) with 5 to 7 terms. Therefore, the expected approximate values of p are: 1<p<5, for ARIMA, it can be considered that there is no trend, i.e. d=0, since it is clear that in PACF under lag 1 there is no value close to +1 or -1, the expected values of q are: 1<q<7. As there is no trend, it can be considered that the series is stationary and there is no trend either towards a reduction, or an increase of PM10 pollution over the examined 5 year period.

In order to determine the most adequate model with the respective values for the parameters (p,d,q), numerous models are constructed with ascending parameter values. When model results are similar, we apply the parsimony principle of choosing the simplest model (Box *et al.*, 1994).

The quality of the models and model fit to data are assessed using the following adequacy tests for time series: the coefficient of determination R^2, root mean square error (RMSE), mean absolute error (MAE), mean absolute percent error (MAPE). Table 2 presents the statistics of two models which provide the best approximation as per these criteria:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y_t - \hat{y}_t\right)^2},$$

$$MAPE = \frac{100}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right|,$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|y_t - \hat{y}_t\right|.$$

where $\hat{y}_t$ is the predicted value at every time t for the transformed variable.

The observed values indicate the percentage of data described by the model which are respectively R^2=0.564 for the model (1,0,5), i.e. the obtained ARIMA describes around 56% of the data and for the model (5,0,7) R^2=0.566. For ARIMA (1,0,5), the autoregression component is (p =1), i.e. the strongest influence on the level of pollution is that of the value for the previous day. The moving average component (q=5) is an indicator that local stochastic changes are correlated with the 5 previous stochastic terms in the time series.

Tables 3 and 4 show that both models have very good statistical significance (Sig.) of the coefficients and the constant considered at level α=0.05.

Table 2. Statistics of ARIMA models for the PM10 pollutant for the city of Pernik.

		ARIMA	**ARIMA**
		(1,0,5)	**(5,0,7)**
Model Fit statistics	Stationary R-squared	0,564	0,566
	R-squared	0,564	0,566
	RMSE	0,086	0,085
	MAPE	3,253	3,247
	MAE	0,063	0,062
	Normalized BIC	-4,889	-4,866
Ljung-Box	Statistics	10,543	4,597
	DF	12	6
	Sig.	0,568	0,960

Table 3. Parameters of the ARIMA (1,0,5) model with the transformed data trPM10.

Parameters		**Estimate**	**SE**	**t**	**Sig.**
Constant		1,969	0,022	88,692	0,000
AR	Lag 1	0,987	0,006	176,236	0,000
MA	Lag 1	0,268	0,024	11,099	0,000
	Lag 2	0,352	0,025	14,269	0,000
	Lag 3	0,110	0,026	4,278	0,000
	Lag 4	0,041	0,025	1,656	0,098
	Lag 5	0,077	0,024	3,246	0,001

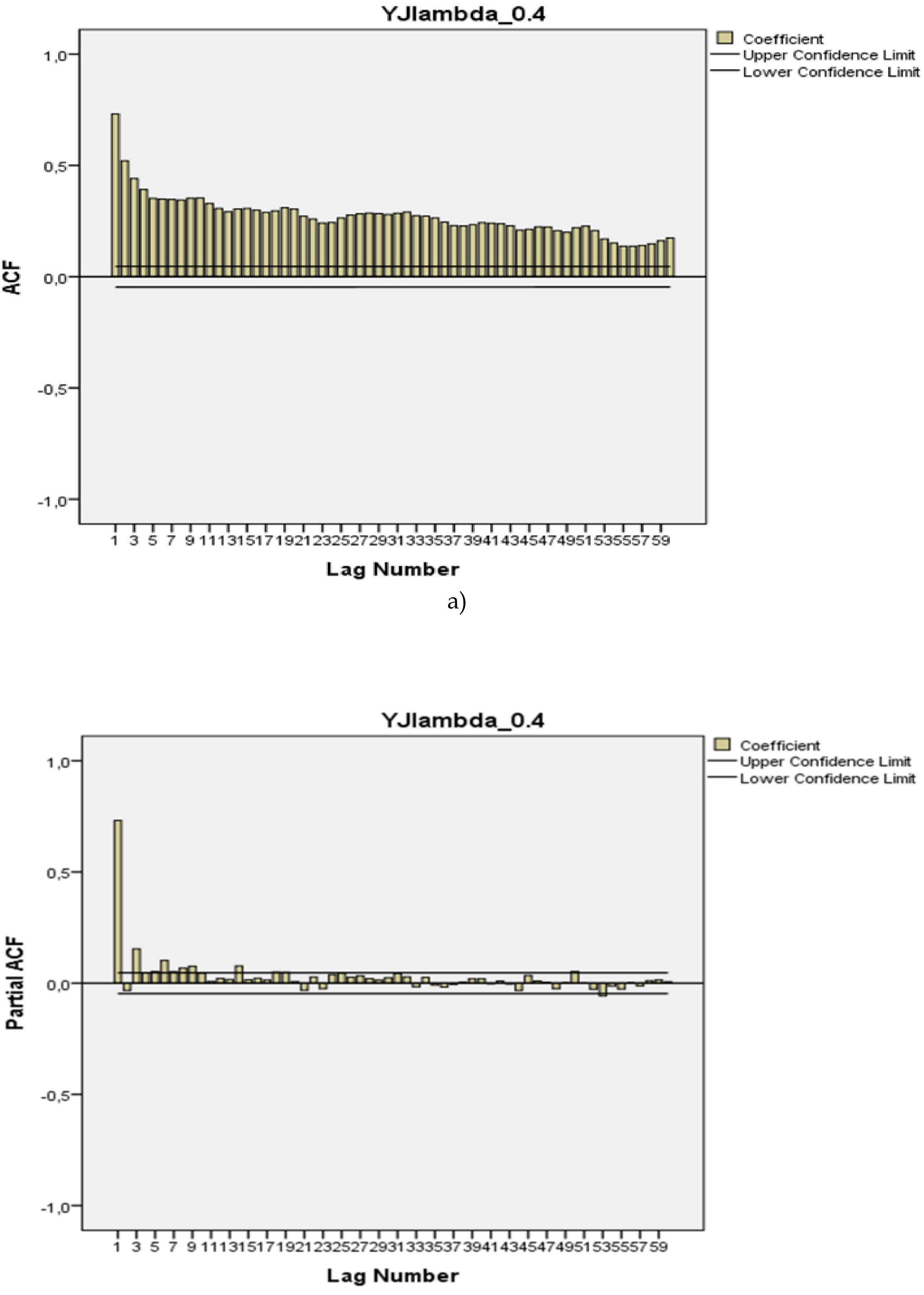
Table 4. Parameters of the ARIMA (5,0,7) model with the transformed data *trPM10*.

Parameters		**Estimate**	**SE**	**t**	**Sig.**
Constant		1,969	0,022	90,176	0,000
AR	Lag 1	1,714	0,237	7,239	0,000
	Lag 2	-2,091	0,504	-4,148	0,000
	Lag 3	1,683	0,595	2,830	0,005
	Lag 4	-0,724	0,453	-1,597	0,110
	Lag 5	0,394	0,179	2,199	0,028
MA	Lag 1	0,998	0,236	4,234	0,000
	Lag 2	-1,220	0,349	-3,501	0,000
	Lag 3	0,553	0,318	1,741	0,082
	Lag 4	-0,042	0,172	-0,242	0,809
	Lag 5	0,196	0,088	2,221	0,027
	Lag 6	0,099	0,056	1,776	0,076
	Lag 7	0,142	0,034	4,226	0,000

The equation in the case of ARIMA (1,0,5) has the form (see also Table 3): trx_t=1.969 + 0.987trx_{t-1} - 0.268a_{t-1} - 0.352a_{t-2} - 0.11a_{t-3} - 0.077a_{t-5}.

In the first model ARIMA (1,0,5), there is one non-significant parameter in lag MA4, which is excluded from the equation, respectively in the second model ARIMA(5,0,7), there are several non-significant parameters, which are also excluded from the equation, respectively with lags AR4 and MA: 3, 4, 6. The statistical indices in Table 2, demonstrate that the two models provide almost the same approximation, but we chose the simpler model (1, 0, 5) to work with.

Fig. 4 illustrates a comparison of the observed values for the PM10 pollutant with those obtained using the ARIMA (1,0,5) model. The actual measured values for PM10 are given in blue, while the predicted values are in green. Very good fit is observed between the actual data and the values predicted by the chosen model.



a)

b)

Fig. 3 a, b. ACF and PCF for PM10 for the city of Pernik.

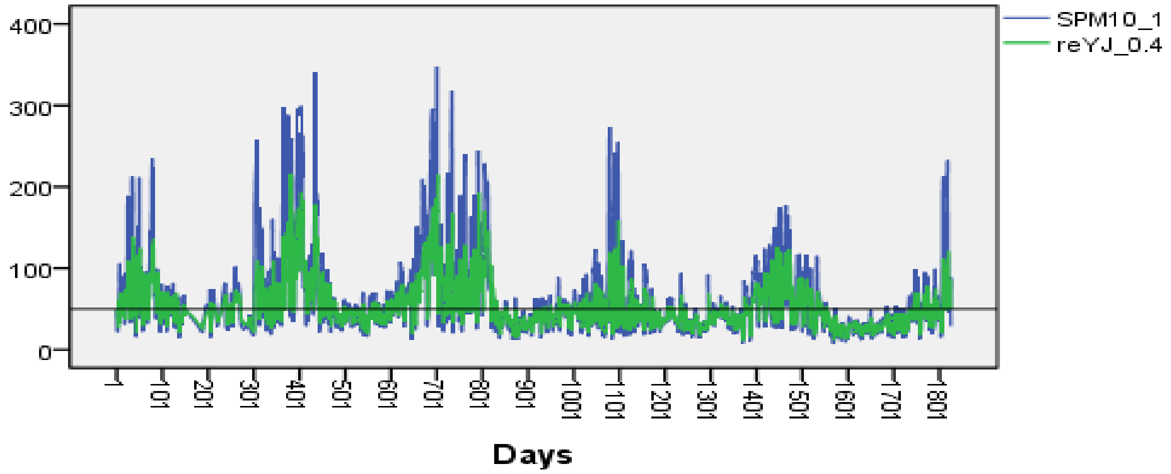


Fig. 4. Comparison of observed values of PM10 against those obtained using the model (1,0,5) following inverse transformation *reYJ*_0.4. The horizontal line indicates the permissible limit of 50 µg/m^3.

Diagnostics of the model

As it was described above the model residuals (the differences between the values predicted by the model and the measured data) are assumed to be random and normally distributed. Fig. 5 shows the distribution of the standardized residuals of the model ARIMA(1,0,5), showing zero mean and standard deviation 1. It is very closed to the standard normal distribution N(0,1).

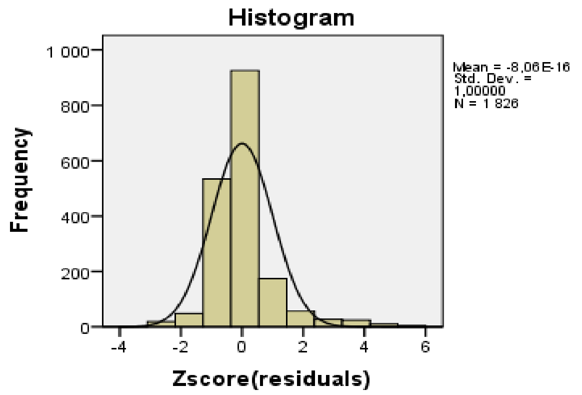


Fig. 5. Distribution of the standardized residuals of the model ARIMA(1,0,5).

Application of the models for forecasting future concentrations

The power of the ARIMA models lies with the good results achieved with predicting future events. In our case, Fig. 6 presents the application of the selected ARIMA(1,0,5) model for short-term forecasting over 7 days – from 25 December to 31 December 2015. To this end, actual additional data is used, which are not included in the construction of the model, and are comparable with the forecasts, obtained using the model. The actual data are colored in green, and the forecast ones in blue. As shown, there is very good fit with the observed variable.

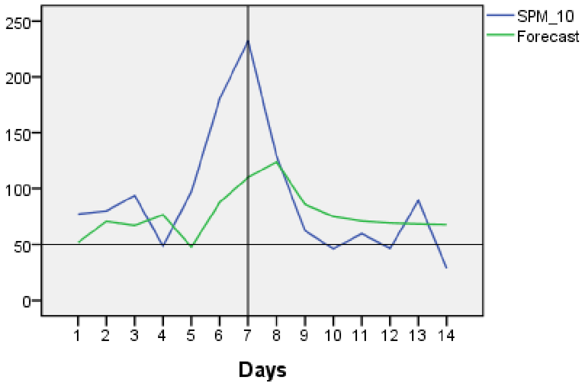


Fig. 6. PM10 values forecast using ARIMA (1,0,5) over a period of 7 days, compared against actual measured values. The horizontal line indicates the permissible limit of 50 µg/m^3 and the vertical line is the limit between the last used 7 days (on the left side) and the forecasted 7 days (on the right side of the line).

Discussion

In order to model PM10 concentrations, univariate stochastic ARIMA models of time series are constructed. It is found that the original series does not contain a trend, which indicates that the chosen 5-year period demonstrates no trends for increasing or decreasing pollutant quantities. At the same time, the values of PM10 are high, and their averages exceed the regulatory requirements. It was found that the chosen modeling approach is adequate. The presented graphic forecast of the model compared against the measured concentration fits well with actual measurements. The model is applied for short-term forecasting of air pollution with PM10 for 7 days ahead. The forecast results indicate bad air quality.

In accordance with the regional reports of Pernik Municipality (RIOSV Pernik, 2015) for the air quality, the main reasons for the high pollution levels of fine particulate matter in the city of Pernik are vehicle traffic and households. During winter, heating in the city is provided mainly by solid fuels – coal, briquettes, and wood; with particulate matter remaining at low altitudes and dissipating only slightly. The existing high levels of the pollutant also result from the construction of the „Lyulin" Motorway until 2012, along with the operations of the steelworks "Stomana Industry", etc. The influence of all air pollutants is further worsened by the city's location within a valley, surrounded by mountains. The specific meteorological conditions such as fog, precipitation, also contribute to the worsening of the problem since they cause the accumulation of pollutants close to the ground over a longer period of time.

Conclusions

This study presents the results from the statistical examination of the PM10 air pollutant in the city of Pernik, west Bulgaria. The data are processed using the ARIMA method for time series analysis. The results obtained by applying this method for PM10 show that the pollutant is problematic and exceeds the permissible limit values. The obtained predictive models provide information about future pollution levels which can be used by the respective authorities to take adequate timely measures in order to reduce the concentrations of PM10.

This is a convenient approach for finding long-term future pollution trends and providing warnings in this regard.

Acknowledgements

This paper is partially supported by Plovdiv University Paisii Hilendarski Scientific Fund – NPD, project SP15-FMIIT-007.

References

BOX G.E.P., G.M. JENKINS, G.S. REINSEL. 1994. *Time series analysis, forecasting and control*, 3rd ed. New Jersey. Prentice-Hall, Inc.

Directive 2008/50/EC. 2008. Directive of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe. - *Official Journal of the European Union*, L 152/1. Available at: [eur-lex.europa.eu]

EEA. 2016. Executive environment agency, National system for real-time air quality control in Bulgaria. *Daily Bulletin for air quality in the country*. Available at: [pdbase.government.bg].

European Commission. 2015. *Air Quality Standards*. Available at: [ec.europa.eu].

EXEA. 2016. Executive Environment Agency, Bulgaria. Available at: [eea.government.bg].

GOCHEVA-ILIEVA S.G., A.V. IVANOV, D.S. VOYNIKOVA, D.T. BOYADZHIEV. 2014. Time series analysis and forecasting for air pollution in small urban area – an SARIMA and factor analysis approach. - *Stochastic Environmental Research and Risk Assessment*, 28 (4): 1045-1060. [DOI].

IBM Corp. 2013. *IBM SPSS Statistics for Windows*, Version 22.0. Armonk, NY: IBM Corp.

IVANOV A., D. VOYNIKOVA, S. GOCHEVA-ILIEVA, H. KULINA, I. ILIEV. 2015. Using principal component analysis and general path seeker regression for investigation of air pollution and CO modeling. – In: Todorov M. (Ed.): *7th Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences.* Melville. American Institute of Physics, AIPCP, vol. 1684(100004) pp. 1-11. [DOI].

IVANOV A., S. G. GOCHEVA-ILIEVA. 2013. Short-time particulate matter PM10 forecasts using predictive modeling techniques. – In: Todorov M. (Ed.): *Fifth Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences.* Melville. American Institute of Physics, AIPCP vol. 1561, pp. 209-218. [DOI].

RIOSV Pernik. 2015. *Report on the state of air quality.* Available at: [pk.riosv-pernik.com].

SHARMA P., A. CHANDRA, S. KAUSHIK. 2009. Forecasts using Box–Jenkins models for the ambient air quality data of Delhi City. - *Environmental Monitoring and Assessment*, 157(1-4): 105-112. [DOI].

SLINI T., A. KAPRARA, K. KARATZAS, N. MOUSSIOPOULOS. 2006. PM10 forecasting for Thessaloniki, Greece. - *Environmental Modelling & Software*, 21(4): 559-565. [DOI].

STADLOBER E., Z. HÜBNEROVÁ, J. MICHÁLEK, M. KOLÁŘ. 2012. Forecasting of daily PM10 concentrations in Brno and Graz by different regression approaches. - *Austrian Journal of Statistics*, 41 (4): 287–310. Available at: [stat.tugraz.at]

VESELY V., J. TONNER, Z. HRDLIVCKOVA, J. MICHALEK, M. KOLAR. 2009. Analysis of PM10 air pollution in Brno based on generalized linear model with strongly rank-deficient design matrix. – *Environmetrics*, 20(6), 676-698. [DOI].

VLACHOGIANNI A., P. KASSOMENOS, A. KARPPINEN, S. KARAKITSIOS, J. KUKKONEN. 2011. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. – *Science of Total Environment*, 409(8): 1559–1571. [DOI].

VOYNIKOVA D. S., S. G. GOCHEVA-ILIEVA, A. V. IVANOV, I. P. ILIEV. 2015. Studying the effect of meteorological factors on the SO2 and PM10 pollution levels with refined versions of the SARIMA model. – In: M. Todorov (Ed.): *7th Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences.* Melville. American Institute of Physics, AIPCP, vol. 1684(100005), pp. 1-12. [DOI].

WHO (World Health Organization). 2013. *Health effects of particulate matter. Policy implications for countries in Eastern Europe, Caucasus and central Asia.* Avaliable at: [euro.who.int].

WHO (World Health Organization). 2015. *Air pollution.* Available at: [who.int].

WILKS D. S. 2011. *Statistical methods in the atmospheric sciences*, 3nd ed. Amsterdam. Elsevier.

YEO I. K., R. A. JOHNSON. 2000. A new family of power transformations to improve normality or symmetry. – *Biometrika*, 87(4): 954–959. [DOI].

ZWOZDZIAK A., L. SAMEK, I. SOWKA, L. FURMAN, M. SKRETOWICZ. 2012. Aerosol pollution from small combustors in a village. - *Scientific World Journal*, Article ID 956401. [DOI].

Received: 27.09.2016
Accepted: 31.10.2016