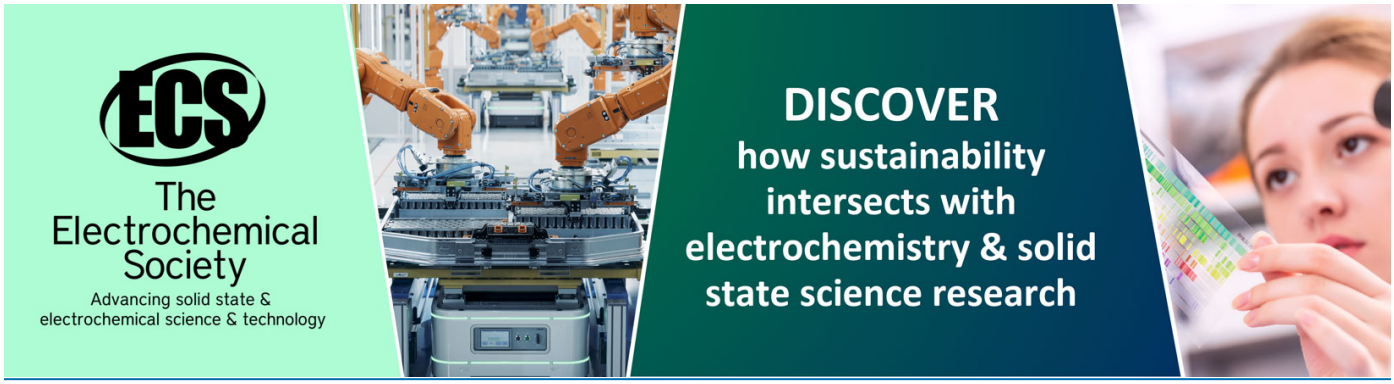
Stochastic modelling of daily air pollution in Burgas, Bulgaria



Stochastic modelling of daily air pollution in Burgas, Bulgaria

S K Koleva, S G Gocheva-Ilieva and H N Kulina

Department of Mathematical Analysis, Faculty of Mathematics and Informatics, University of Plovdiv Paisii Hilendarski 24 Tzar Asen Street, 4000 Plovdiv, Bulgaria
E-mails: stkoleva@uni-plovdiv.bg, snow@uni-plovdiv.bg, kulina@uni-plovdiv.bg

Corresponding author's e-mail address: stkoleva@uni-plovdiv.bg

Abstract. Exceeding the norms and limits of atmospheric air pollution causes enormous damage to the population's health and the environment. Determining the factors affecting air quality is a current task in a local, regional, and global scale. In this study, we use daily time series data for the main air pollutants in Burgas, Bulgaria – O_3, NO, NO_2, CO, SO_2, and PM_{10}, to analyze, model, and forecast these levels depending on meteorological factors. For this purpose, the stochastic ARIMA method and ARIMA with transfer functions are applied. Results are obtained for univariate and multivariate time series. Particular attention is paid to the concentrations of the secondary pollutant ground-level ozone (O_3), which are modelled as a function of all variables considered. Results were evaluated using root mean square error, mean absolute percentage errors, and the coefficient of determination. Short-term forecasts have been obtained for seven days ahead. Model accuracy up to 84% has been established.

1. Introduction

The quality of the air we breathe and its pollution are among the most pressing topics to date. There are 36 automated and certified air control stations in Bulgaria and the main air pollutants in the larger cities of the country are monitored systematically. For the whole country, monitoring is carried out by the Executive Environment Agency, European and national norms and standards, and World Health Organization documents [1-3]. Although in recent years, harmful emissions into the ambient air have decreased, some systematic exceedances of air pollution for several cities in Bulgaria are still observed. It should be clarified that air pollution for each region depends to a significant extent not only on its weathering and geographical and climatic characteristics but also on a large number of anthropogenic factors. Such factors are the results of human activity, including various production processes, road traffic, emissions of harmful emissions from households, and others. The characteristics and harmful effects of air pollutants on human health are described, for example, in [1-3].

The availability of a huge volume of empirical measurement data in the field of air pollution enables their analysis and statistical modelling to derive trends and dependences, as well as to predict the future state of ambient air quality. A large number of studies on data processing on atmospheric pollutants have been published in the scientific literature. At the same time, in Bulgaria, such studies are relatively few. To analyse and model concentrations of the main air pollutants, such as O_3 – ground-level ozone, CO – carbon monoxide, NO – nitrogen monoxide, SO_2 – sulfur dioxide, NO_2 – nitrogen dioxide, PM_{10} – fine particulate matter below 10 microns and others different statistical and numerical methods were applied. The classical linear stochastic ARIMA (Auto Regressive Integrated Moving Average) approach and its variants [4] have been used in [5-9]. In [5], a stationary stochastic ARMA/ARIMA modelling approach has been considered to forecast the daily mean air pollutants O_3, CO, NO, and NO_2

concentration. The models were calibrated by means of different information criteria, such as the Akaike Information Criterion (AIC), Hannon–Quinn Information Criterion (HIC), Bayesian Information criterion (BIC), etc. and the plots of the Autocorrelation function (ACF), and Partial autocorrelation function (PACF). Guarnaccia et al. [6] built univariate seasonal ARIMA (SARIMA) models for predicting CO concentration levels in a district of Monterrey, Mexico and concluded that the use of hourly data allowed the achievement of reliable prediction especially on a short time of 24-hour period. The authors of [7] apply ARIMA to model, analyse and forecast SO_2 and NO_2 time series values, using average daily measurements. Statistical studies on PM_{10} pollution in two Bulgarian cities - Pernik and Ruse, are presented in [8,9]. A two-step SARIMA approach was developed in [10] for PM_{10} concentration prediction. To achieve a real-world effect, meteorological variables are first predicted with one-dimensional SARIMA models, then the pollutant data is modelled depending on the historical and newly calculated predictive values of the predictors.

In recent years, cutting-edge approaches based on machine learning (ML) algorithms have been actively used to model and predict empirical ambient air data. Many authors combine or compare their results with ARIMA and other classical statistical approaches, hybrid models are also developed. Wavelet analysis with energy spectrograms, combined with the multilayer feed-forward back propagation NN algorithm and compared with ARIMA are reported in [11] to forecast PM_{10} monthly levels in selected regions of Chennai. Discrete Haar wavelet transform and ARIMA models were coupled to improve ozone prediction in [12]. Principal Component Analysis (PCA) and path seeker technique are implemented in [13] to model the CO concentration. Paper [14] deals with the Random Forest algorithm for analysis and forecasting of PM_{10} pollution. The authors of [15] construct hybrid Elman ANN and ARIMA models using empirical data of SO_2. Boosted trees method with regularized regression from ensemble learning group are applied for studying ground-level ozone and PM_{10} in [16]. In [17] the authors develop several ML models of O_3 and PM_{10}, based on hourly data. The best performance is obtained using the stochastic Gradient Boosting method, representative of ensemble tree-based methods, demonstrating computational efficiency and robustness to overfitting. More results on ML methods and case studies in the field of air pollutants could be found in the review article [18].

The aim of this study, based on the collected empirical data for the city of Burgas are as follows: 1) To apply univariate ARIMA methods for stochastic regression modelling and forecasting separately the concentrations of the six air pollutants O_3, SO_2, NO, NO_2, CO and PM_{10}; 2) To create multivariate ARIMA models of each of these time series, depending on meteorological time series; 3) To construct the multivariate ARIMA model of tropospheric ozone; 4) To evaluate the models' performance and their residuals; 5) To validate the models by short-term predictions of pollution.

The study was carried out using IBM SPSS Statistics [19,20].

2. Data and methods

2.1. Description of data and initial data processing

This section presents the data and the results from the initial analysis performed on them.

2.1.1. Study area

Burgas is the second largest city on the Bulgarian Black Sea coast. Here is the largest chemical and oil refinery in Southeastern Europe, which is also the largest employer in Bulgaria, the international and second busiest Bulgarian airport, as well as the most significant Bulgarian port and the only oil port in the country. The city is located on the westernmost point of the Black Sea. The climate is humid subtropical with wet and continental influence. The environment area is urbanized, with high building density, intensive car traffic and industrial activity in the municipality of Burgas. Three estuarine lakes are located on the territory of the city. A specific influence is also exerted by the sea breeze, whose circulation has a direct impact on the climate and the dispersion of atmospheric pollutants.

2.1.2. Data

In this article, we explore data on six major air pollutants for the city of Burgas for the period from the 1 January 2019 to 31 March 2023 with average daily data or for N=1551 days. In particular, we are looking for mathematical models describing the behaviour of O_3, μg/m^3 (ground level ozone concentrations), CO, mg/m^3 (carbon monoxide), NO, μg/m^3 (nitrogen monoxide), SO_2, μg/m^3 (sulfur dioxide), NO_2, μg/m^3 (nitrogen dioxide), PM_{10}, μg/m^3 (fine particulate matter with diameter below 10 microns) in relation to meteorological data.

The following seven meteorological variables are used to build the multivariate models: MaxT, °C (maximum daily mean air temperature), MinT, °C (minimum daily average air temperature), Cloud, % (cloudiness), Humidity, % (relative air humidity), Precipi, mm (precipitation), Speed, m/s (wind speed), Pressure, mbar (ambient air pressure). The data were recorded from the automated measured station situated in Dolno Ezerovo, a corner of the city of Burgas. The GPS coordinates of the station Dolno Ezerovo are 42°31′05.99" N 27°22′22" E. Data were retrieved from the official sites [21,22].

The maximum number of missing values for each initial time series is less than 2%. In the analyses the missing values were replaced using linear interpolation. Several extreme values of pollutants were found and replaced by the next largest. The basic descriptive statistics of time series of pollutants and meteorological variables are given in Table 1.

Table 1. Descriptive statistics of the examined data of air pollutants and meteorological data for the city of Burgas.

Variable	Mean	Median	Skewness	Kurtosis
O3, μg/m^3	44.52	45.88	-0.14	-0.59
NO, μg/m^3	4.90	4.94	1.38	4.28
NO2, μg/m^3	14.10	13.17	0.72	0.45
CO, mg/m^3	0.31	0.25	1.48	2.89
SO2, μg/m^3	10.85	10.63	0.66	1.51
PM10, μg/m^3	33.79	30.21	1.47	3.12
MaxT, °C	17.59	17.00	0.01	-1.05
MinT, °C	10.49	10.00	-0.08	-0.98
Cloud, %	0.38	0.32	0.57	-0.86
Humidity, %	0.71	0.71	-0.57	1.77
Precipi, mm	1.70	0	5.01	31.48
Speed, m/s	1.47	1.31	1.56	3.80
Pressure, mbar	1010.3	1010.6	-33.62	1256.29

The next Figure 1 shows the sequences plots of the initial time series of the examined air pollutants. There are pronounced seasonal changes. Higher concentrations of pollution for SO_2, NO, NO_2, CO and PM_{10}, are observed in winter period while in summer the levels drop. This behaviour is opposite for ozone. This indicates that weather conditions have a strong influence on pollutants' emissions. The remaining causes of increased concentrations of harmful emissions, including car traffic, seasonal combustion processes from households and manufacturing enterprises, and others, are relatively constant and do not change sharply over time, compared to meteorological changes. Their influence is stochastically comprised in the PM_{10} values. This will make it easier for us to create models to predict pollutant concentrations for a short period of time ahead based on the meteorological conditions.

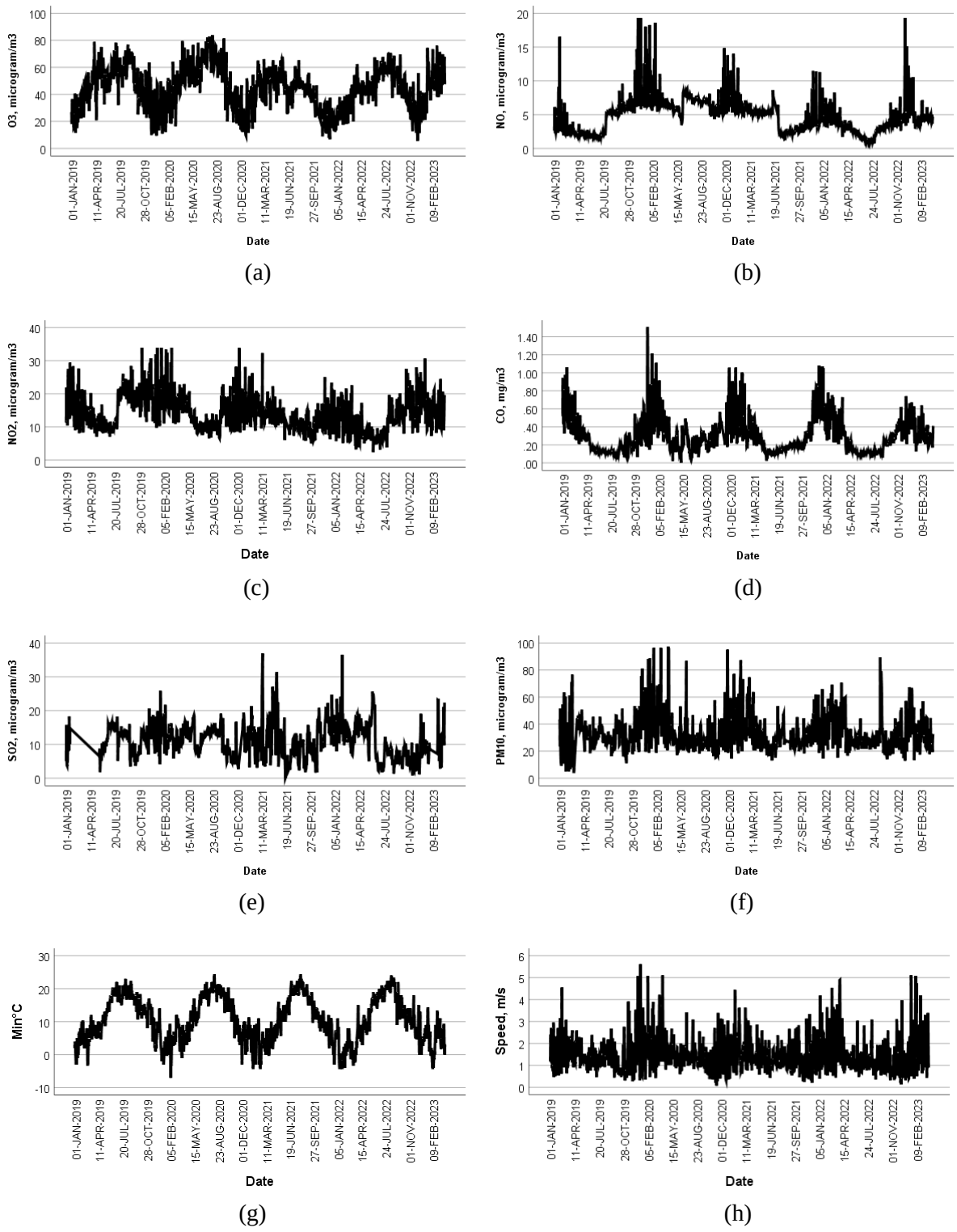


Figure 1. Sequence plots of the investigated variables: (a) O3, (b) NO, (c) NO2, (d) CO, (e) SO2, (f) PM10, (g) Tmin, (h) Speed.

2.1.3. Methods used

We will apply the Box-Jenkins ARIMA methodology [4]. The general form of the ARIMA model of a given time series variable Y is represented as ARIMA (p, d, q) where p is the number of autoregressive terms, d is the number of differences, q is the number of moving average terms. The main requirements for building an adequate ARIMA model include normal or close to normal distribution, linear dependence, absence of missing values and stationarity [4,23].

- Univariate ARIMA

When the time series is stationary, i.e. the probability distribution does not depend on time, so its mean and variance are constant, the univariate AR process of order p is described by a difference equation of order p of the type:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + a_t = \left(\sum_{j=1}^{p} \phi_j B^j \right) Y_t + a_t, \quad t = p+1,...,N \tag{1}$$

where $(\phi_1, \phi_2, ..., \phi_p)$ are constant coefficients (model parameters), a_t. is a stochastic term (random error, white noise) for each time t, under the assumption $a_t \sim WN\left(0, \sigma^2\right)$, B is the backward (lag) operator $BY_t = Y_{t-1}$.

If the process contains mostly systematic random fluctuations a_t. around some fixed level, then it is defined as a stochastic MA process. The MA model with moving averages is of order q if, for each t, it is represented by a difference equation of order q of the type

$$Y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} = \left(1 - \sum_{j=1}^{q} \theta_j B^j \right) a_t, \quad t = q+1,...,N \tag{2}$$

where $\theta_1, \theta_2, ..., \theta_q$ are constant parameters.

In the general case, ARIMA (p,d,q) is written as

$$\left(1 - \sum_{j=1}^{p} \phi_j B^j \right) (1-B)^d Y_t = \left(1 - \sum_{j=1}^{q} \theta_j B^j \right) a_t + c\,. \tag{3}$$

In (3) c is a constant,

- Multivariate ARIMA

When the dependent time series Y is considered according to other time series $X_{1t}, X_{2t}, ..., X_{kt}$ which values are known in the same time period, a multidimensional ARIMA could be used. It is called ARIMA with transfer functions (TF) [4]. The ARIMA/TF model is written as:

$$(1-B)^d Y_t = \frac{MA}{AR} a_t + \sum_{i=1}^{k} \left(\frac{Num_i}{Den_i} (1-B)_i^{d_i} B^{b_i} X_{it} \right) + \mu \tag{4}$$

where μ is constant, B^{b_i} is a delay member with a positive integer lag b_i , MA and AR, Num_i and Den_i have the form of the difference polynomials with constant coefficients:

$$AR = \left(1 - \sum_{j=1}^{p} \phi_j B^j \right), \; MA = \left(1 - \sum_{j=1}^{q} \theta_j B^j \right),$$

$$Num_i = \left(\sum_{j=0}^{u} \omega_{ij} B^j \right), \; Den_i = \left(1 - \sum_{j=1}^{v} \gamma_{ij} B^j \right) \tag{5}$$

To obtain statistically correct ARIMA or ARIMA/TF models, the significance of its coefficients must be achieved at a set level α (usually 0.05).

2.1.4. Data transformation

The basic assumptions for ARIMA analysis include the normal or close to normal distribution and stationarity of the time series involved [23]. As it is seen from Table 1, the coefficients of skewness and kurtosis of the considered time series are different from zero, therefore the normality condition is violated. To stabilize the variance and improve the distribution to normality the Yeo-Johnson power transformation was used by the expressions [24]:

$$trY = YJ(\lambda, Y) = \begin{cases} \left\{(Y+1)^{\lambda} - 1\right\} / \lambda & Y \geq 0,\ \lambda \neq 0 \\ \log(Y+1) & Y \geq 0,\ \lambda = 0 \\ -\left\{(-Y+1)^{2-\lambda} - 1\right\} / (2-\lambda) & Y < 0,\ \lambda \neq 2 \\ -\log(-Y+1) & Y < 0,\ \lambda = 2 \end{cases}, \qquad \lambda \in [-2, 2] \tag{6}$$

where Y is the original variable, trY is its transformed variable, and λ is a real parameter. The parameter λ is tuned by a stepwise procedure using some statistical tools or normality test.

Table 2 shows the obtained values for λ, and Skewness and Kurtosis of the transformed pollutant's variables. With the small values of Skewness and Kurtosis, it can be concluded that the distributions of the transformed variables are close to a normal distribution. The box plots of the standardized transformed variables are presented in Figure 2. It is observed the improvement of the distribution, however, the outliers exist, and could not be neglected.

Table 2. Statistics of the transformed variables of air pollutants.

Transformed pollutant variable	Parameter λ	Skewness	Kurtosis
trO3	1.2	0.018	-0.600
trNO	0.6	0.665	1.248
trNO2	0.2	-0.014	-0.201
trCO	-1.8	0.437	-0.564
trSO2	0.8	0.352	0.708
trPM10	-0.2	-0.429	2.724

2.1.5. Statistical measures to assess the quality of models

The quality of models is evaluated by standard statistical indicators: coefficient of determination, root mean square error, mean absolute percentage error (R^2, RMSE, MAPE), calculated by the expressions:

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

where $\hat{Y}_t$ is the value predicted by the model at time t, $\bar{Y}$ is the mean value of Y, N is the sample size. We will look for models for which the coefficient of determination is close to 1 and the errors are small.

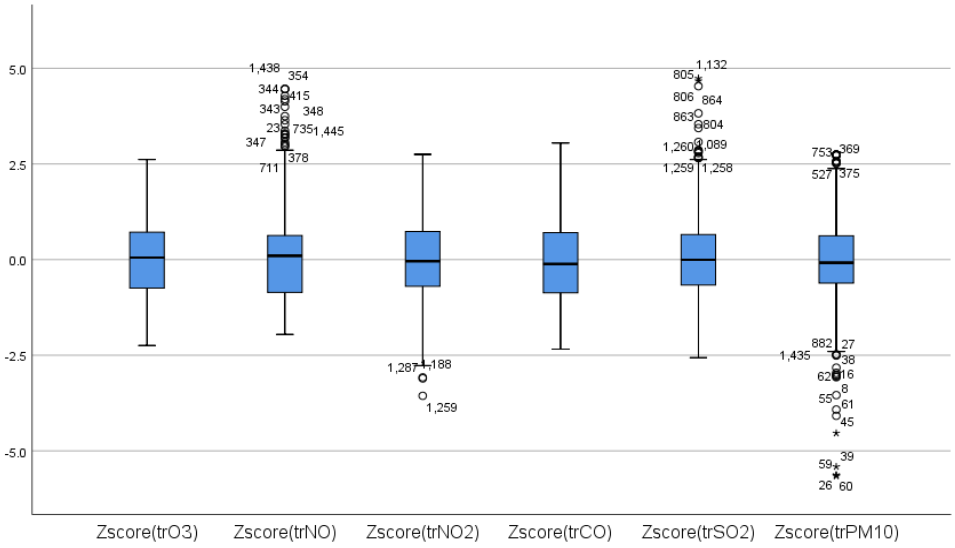


Figure 2. Box plots of the standardized transformed variables of air pollutants.

3. Results and discussion

Using the developed methodology, univariate and multivariate models of transformed air pollutant variables have been constructed. Univariate models are a special case when there are no predictors.

All models were built using data for N=1544 days. The data of the last 7 days will be used for model validation and forecasting the pollution level.

3.1. Building and evaluating the univariate ARIMA models

For each variable of the six pollutants, separate univariate ARIMA models were built using their corresponding transformed variables. The process of constructing an ARIMA model requires five basic steps to correctly determine the parameters p, d, q.

1) The corresponding sequence plots of the ACF and PACF are constructed and examined to reveal the presence of trends (*d*). The lags outside the permissible values are inspected to find initial values of *p* and *q*.

2) In the presence of a PACF coefficient with a value close to 1 or -1, an augmented Dickey-Fully (ADF) test is conducted to specify whether it is a trend or not and to determine the exact value of the parameter *d*.

3) An ARIMA (p,d,q) is constructed with selected parameters.

4) The statistical significance of the parameters is checked. In our case, the significance level should be Sig.<0.05. Insignificant parameters are removed and the values of *p*, *q* are adjusted. The parameter selection procedure can be repeated many times.

5) A detailed analysis of the residuals of the model is carried out to establish its statistical adequacy. The residual values of the model are investigated to be within the appropriate confidence interval using their autocorrelation functions (ACF). It is also recommended to check the residuals with the Ljung-Box portmanteau test or other appropriate statistical tests. For this purpose, such a test must be insignificant, i.e., Ljung-Box Sig. >0.05. This means that the residuals of the ARIMA model do not contain autocorrelation.

When obtaining several different adequate ARIMA models, the simpler one is chosen or the one that has the best statistical indicators, in this case the attention is on the indicators of (7).

The next Figure 3 shows the univariate ARIMA predictions for transformed variables.

Table 3. Statistics of univariate models of transformed variables of air pollutants.

Transformed variable	Univariate ARIMA model	R^2	RMSE	MAPE	Ljung-Box Sig.
trO3	ARIMA(1,0,6)	0.756	16.442	20.610	0.058
trNO	ARIMA(1,0,5)	0.775	0.577	10.842	0.051
trNO2	ARIMA(1,0,10)	0.611	0.376	8.370	0.314
trCO	ARIMA(1,0,21)	0.812	0.036	16.602	0.210
trSO2	ARIMA(1,0,13)	0.715	1.582	20.163	0.246
trPM10	ARIMA(1,0,3)	0.477	0.131	3.819	0.533

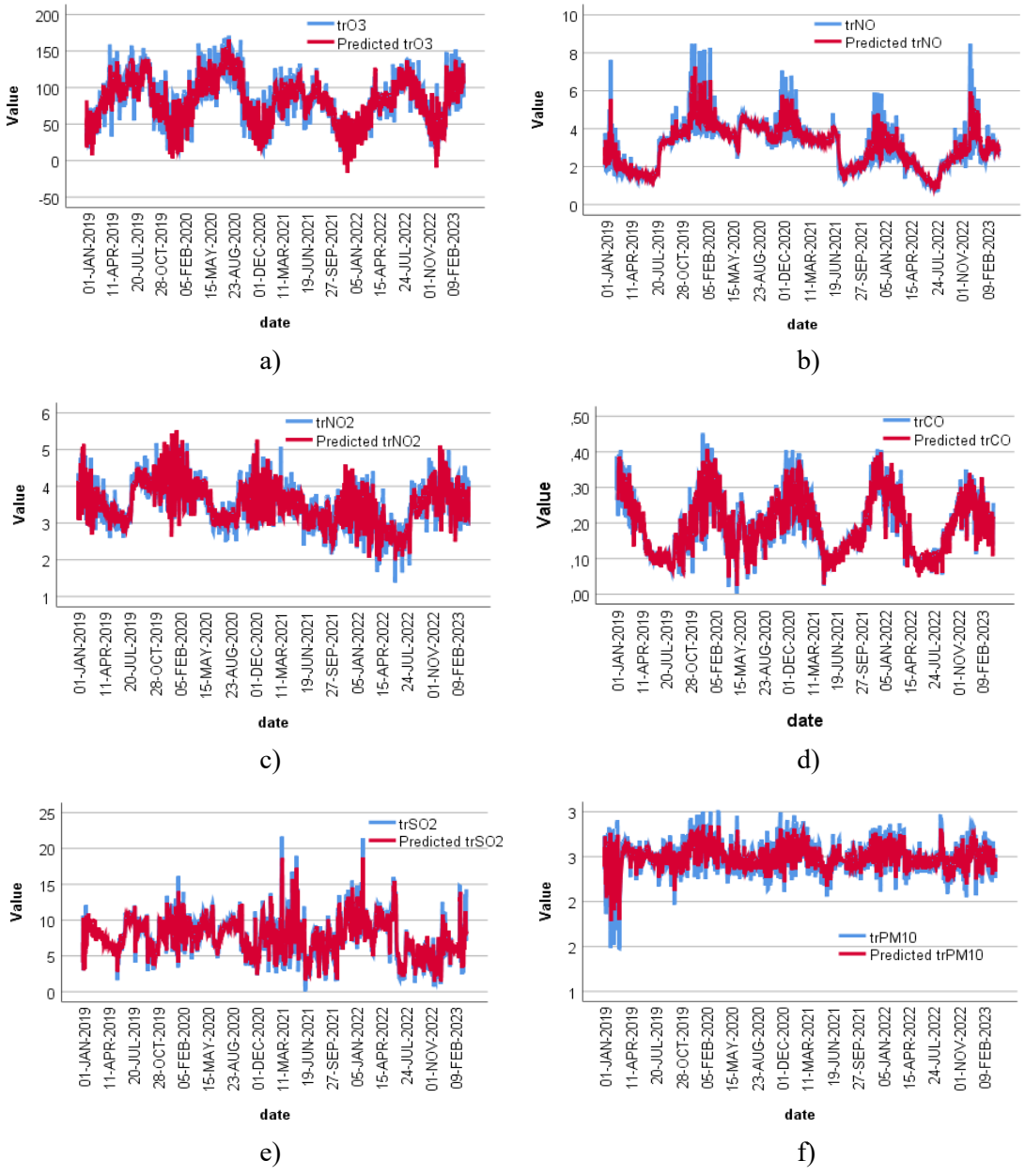


Figure 3. Line plots of the transformed versus modelled values of the six studied air pollutants.

3.2. Building the multivariate ARIMA models

ARIMA/TF models are built for each pollutant separately. The seven meteorological time series from Table 1 were used as predictors. The basic statistics of the selected models are given in Table 4. The comparison with the statistics of the univariate models shows that all models with transfer functions have improved values over the univariate models - both with a larger data matching rate (R^2) and lower RMSE values, as well as the MAPE drops.

Table 4. Summary statistics of the built air pollutant's ARIMA/TF models (for transformed variables).

Transformed dependent variable	ARIMA model	R^2	RMSE	MAPE	Ljung-Box Sig.	Predictors
trO3	ARIMA/TF (1,0,4)	0.840	13.356	16.004	0.368	MinT, Speed, Humidity, Precipi, trNO, trNO2, trCO, trSO2, trPM10
trCO	ARIMA/TF(1,0,3)	0.854	0.032	15.853	0.178	MinT, Speed, Cloud, Pressure
trNO	ARIMA/TF(1,0,5)	0.808	0.535	10.766	0.096	MaxT, Cloud, Humidity, Pressure
trSO2	ARIMA/TF(1,0,13)	0.726	1.552	19.774	0.359	MaxT, Pressure, Humidity
trNO2	ARIMA/TF(1,0,10)	0.749	0.303	6.693	0.097	MinT, Speed, Cloud, Pressure
trPM10	ARIMA/TF(1,0,3)	0.531	0.124	3.606	0.522	MaxT, MinT, Speed, Cloud

3.3. Study on the residuals of the multivariate ARIMA model of ozone

We have paid particular attention to the secondary air pollutant ground-level ozone which was modelled as a function of all the variables considered. 84% coefficient of determination was achieved. In this subsection, we present the results of the residuals diagnostics of the model ARIMA/TF (1,0,4) of trO3. Figure 4 illustrates the ACF and PACF of the residuals of the ARIMA/TF ozone model for 24 lags. It snows a lack of serial correlation. Figure 5 shows the box plot of the residuals and Figure 6 shows the histogram of the residuals of this model. We will add that the formal Kolmogorov-Smirnov test to check the normality of the residuals is small with a statistic equal to 0.035, df=1544, and is significant due to the presence of outliers. In general, we could conclude that the distribution of the residuals is close to normal.

The model performance in terms of the coefficient of determination R^2 is illustrated in Figure 7.

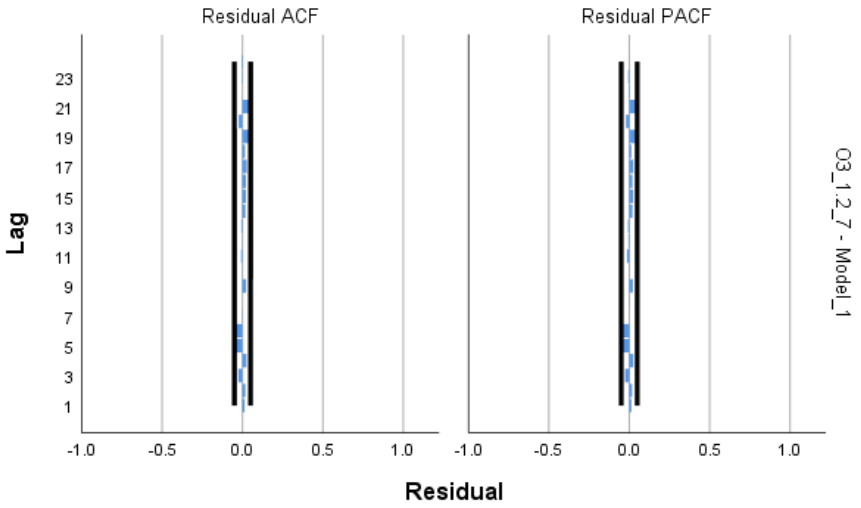


Figure 4. ACF and PACF figures of the residuals of the ARIMA/TF model of trO_3_7.

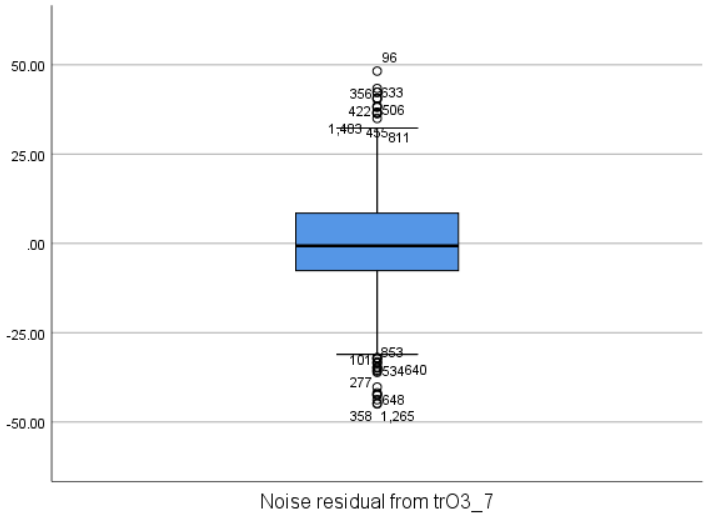


Figure 5. Box plot of model residuals for ozone.

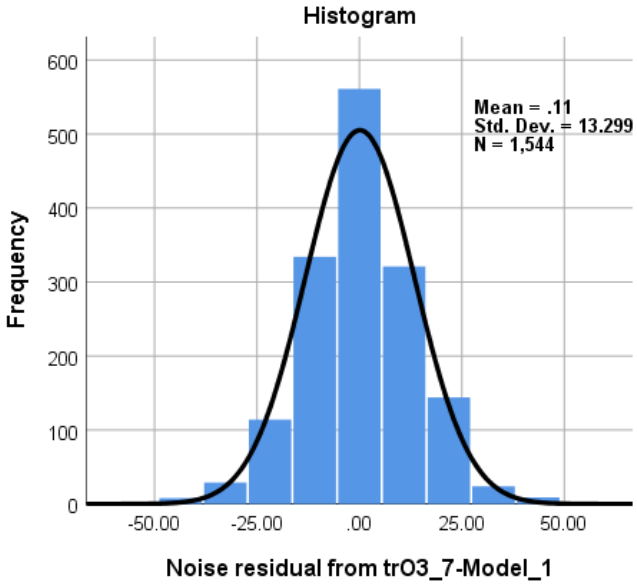


Figure 6. Histogram of ARIMA/TF ozone model residuals

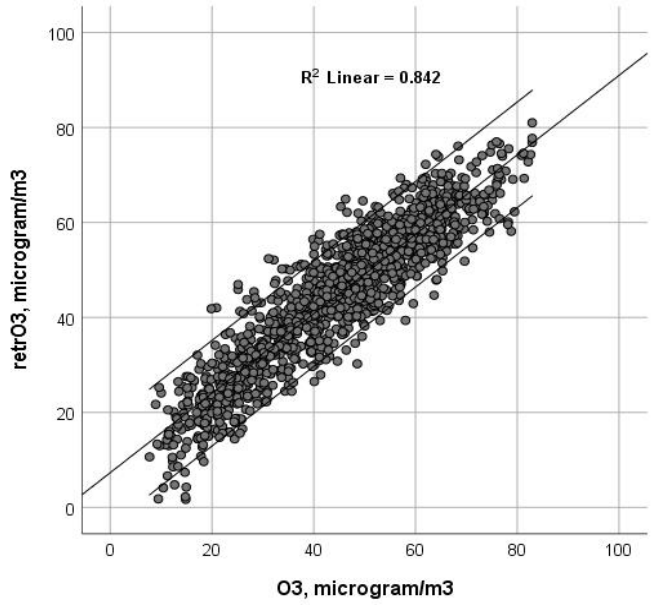


Figure 7. Scatterplot for achieved R^2=84% model fit to data obtained by the retransformed predicted values from the ozone model with a 95% confidence interval.

3.4. Application of ARIMA/TF models to predict future concentrations

To validate the model, from all available data of the dependent variable (trO3) the last few values could be deleted. In our case, these are the data for the last 7 days. To obtain the forecasts, it is necessary to know the values of the predictors (i.e. meteorological variables and the other pollutants). In the real case, the synoptic weather forecasts could be used and the required values of the pollutant variables must be predicted. The O_3 forecasts were obtained without the participation of the last 7 ozone values in the modeling procedure. In Figure 8. a comparison is shown of the observations and model ARIMA/TF

(1,0,4) predictions for the previous 7 days (on the left side of the vertical line) and the observations and forecasted pollutant data for the validated last 7 days (on the right of the vertical line).

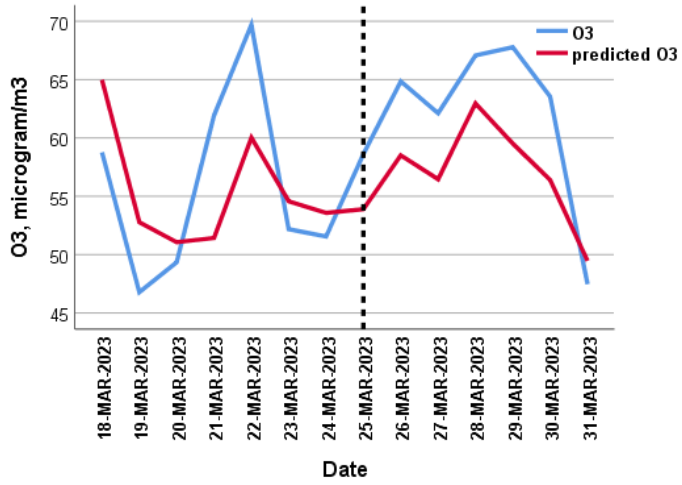


Figure 8. Observed versus predicted ozone values using the ARIMA/TF model for the last 7 days (March 18 to 24, 2023, to the left of the vertical line) and observed versus forecasts using the ARIMA/TF model for the last 7 days (March 25 to 31, 2023, to the right of the vertical line).

The conducted analysis and diagnostics of the model residuals in this subsection give us reason to conclude that the constructed ARIMA/TF(1,0,4) ozone model (see Table 4) is statistically valid and capable of predicting concentrations in a short period ahead.

4. Conclusion

In this study, we investigated average daily data for six air pollutants and seven meteorological variables for the city of Burgas, Bulgaria over a period of 4 years and 3 months. The Box-Jenkins stochastic method for time series analysis was applied for modelling. Univariate ARIMA and multivariate ARIMA/transfer function models were built and statistically investigated. In the preprocessing stage, we used the Yeo-Johnson transformation to stabilize variance and improve the distribution of data. Applying the ARIMA/TF allowed to build models reviling the impact of the given 7 meteorological variables on the considered 6 air pollutants separately. Statistical analyses were carry out to evaluate the model performance and conduct the analyses of the model's residuals, thus validated the model adequacy and usefulness. Especially, the ozone model was built and 84% of data fit was achieved. The model was applied for 7-days ahead forecasting.

More sophisticated modeling approaches are planned as future work, such as ML, discrete wavelet transform, singular spectrum analysis, or others to improve the performance of models and forecasts.

Acknowledgement

This study was supported by the Bulgarian National Science Fund, Grant KP-06-N52/9.

References

[1] Executive Environment Agency, Bulgaria, http://eea.government.bg/bg/soer/2014/air/kachestvo-na-atmosferniya-vazduh
[2] European Environment Agency, https://www.eea.europa.eu//publications/status-of-air-quality-in-Europe-2022
[3] Air Quality Standards. European Commission. Environment, http://ec.europa.eu/environment/air/quality/standards.htm

[4] Box G E P, Jenkins G M and Reinsel G S 1994 *Time Series Analysis, Forecasting and Control*, 3rd edn (New Jersey: Prentice-Hall, Inc)
[5] Kumar U and Jain V K 2010 ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO) *Stochastic Environmental Research and Risk Assessment* **24**(5) 751–760
[6] Guarnaccia C, Breton J G C, Breton R M C, Tepedino C, Quartieri J and Mastorakis N E 2018 ARIMA models application to air pollution data in Monterrey, Mexico *AIP Conference* Proceedings **1982**(1) 020041
[7] Gourav, Rekhi J K, Nagrath P and Jain R 2020 Forecasting air quality of delhi using arima model *Lecture Notes in Electrical Engineering* **612** 315–325 (Singapore: Springer)
[8] Stoimenova M P 2016 Stochastic modeling of problematic air pollution with particulate matter in the city of Pernik, Bulgaria *Ecologia Balkanica* **8**(2) 33–41
[9] Veleva E, Filipova M and Zheleva I 2022 Statistical study of particulate matter (PM10) air contamination in the city of Vidin, Bulgaria *AIP Conference Proceedings* **2522**(1) 050014-1–050014-11
[10] Gocheva-Ilieva S, Ivanov A and Iliev I 2019 Exploring key air pollutants and forecasting particulate matter PM10 by a two-step SARIMA approach *AIP Conference Proceedings* **2106** 020004
[11] Angelena J P, Raj A S, Viswanath J and Muthuraj D 2021 Evaluation and forecasting of PM10 air pollution in Chennai district using Wavelets, ARIMA, and Neural Networks algorithms *Pollution* **7**(1) 55–72
[12] Salazar L, Nicolis O, Ruggeri F, Kisel'ák J and Stehlík M 2019 Predicting hourly ozone concentrations using wavelets and ARIMA models *Neural Computing and Applications* **31** 4331–4340
[13] Ivanov A, Voynikova D, Gocheva-Ilieva S, Kulina H and Iliev I 2015 Using principal component analysis and general path seeker regression for investigation of air pollution and CO modeling *AIP Conference Proceedings* **1684** 1–11
[14] Ivanov A, Gocheva-Ilieva S and Stoimenova-Minova M 2022 Random forest regression for statistical modeling and forecasting of PM10 *AIP Conference Proceedings* **2522**(1) 100005
[15] Sánchez A B, Ordóñez C, Lasheras F S, de Cos Juez F J and Roca-Pardiñas J 2013 Forecasting SO2 pollution incidents by means of Elman artificial neural networks and ARIMA models *Abstract and Applied Analysis* **2013** Art ID 238259
[16] Ivanov A, Gocheva-Ilieva S and Stoimenova M 2020 Hybrid boosted trees and regularized regression for studying ground ozone and PM10 concentrations *AIP Conference Proceedings* **2302** 060005
[17] Krylova M and Okhrin Y 2022 Managing air quality: Predicting exceedances of legal limits for PM10 and O3 concentration using machine learning methods *Environmetrics* **33**(2) e2707
[18] Bai L, Wang J, Ma X and Lu H 2018 Air pollution forecasts: an overview *Internacional Journal of Environmental Research and Public Health* **15**(4) 780
[19] IBM SPSS Statistics, https://www.ibm.com/products/spss-statistics
[20] Yordanova L, Kiryakova G, Veleva P, Angelova N and Yordanova A 2021 Criteria for selection of statistical data processing software *Proceedings of the IOP Conference Series: Materials Science and Engineering* **1031**(1) 012067
[21] Regional Inspectorate of Environment and Water – Burgas, https://www.eea.government.bg/kav/reports/air/qReport/92/01#
[22] World Weather API and Weather Forecast, https://www.worldweatheronline.com/
[23] Wilks D S 2011 *Statistical Methods in the Atmospheric Sciences* 3rd edn (Amsterdam: Elsevier)
[24] Yeo I K and Johnson R A 2000 A new family of power transformations to improve normality or symmetry *Biometrika* **87**(4) 954–959