TIME SERIES ANALYSIS AND PREDICTION OF AIR POLLUTANT CONCENTRATIONS (PM_{10}, SO_2 AND NO_2) USING LINEAR REGRESSION APPROACH IN KAZANLAK, BULGARIA

Svetla Stoykova[1], Miroslava Ivanova[2], Diyana Dermendzhieva[1], Lilko Dospatliev[3]*

[1] Departrment of Applied Ecology and Animal Hygiene, Trakia University, 6000 Stara Zagora, Bulgaria

[2] Department of Informatics and Mathematics, Faculty of Economics, Trakia University, 6000 Stara Zagora, Bulgaria

[3] Departrment of Pharmacology, Physiology and Animal Physiological Chemistry, Trakia University, 6000 Stara Zagora, Bulgaria

Abstract

The aim of the present study is to conduct time series analysis and to predict of air pollutant concentrations applying linear regression approach using PM_{10}, SO_2 and NO_2 data from January, 01, 2023 to April 30, 2024 in the second largest city in the region Stara Zagora, Bulgaria - Kazanlak. All statistical computing, test and graphics were performed with the statistical software R. We received three linear regression models, which showed that if SO_2 increase by 1%, the effect of this increase would result in an increase in NO_2 by 0.90%; if PM_{10} increase by 1%, the effect of this increase would result in an increase in SO_2 and NO_2 by 0.19% and 0.23%, respectively.

Key words: *air pollution, linear regression model, statistical software R, Bulgaria*

Introduction

Air pollution represents an important environmental issue in worldwide since the 70s of the last century [1]. Recognizing the human health risk, the World Health Organization (WHO) has been developing air quality standards and goals. However, despite all global efforts, a survey showed that no country met the WHO air quality standards in 2021 [2].

Over the past decades', many efforts have been made in the development of data science software tools and programming languages allowing the integration of mathematics models of data analysis, visualization, and forecasting [3]. An example of such software is R, used in research such as [4]. One of the used efficient model of time series analysis of air pollutants is linear regression applied in articles such as [5].

In this context, the aim of the present study is to conduct time series analysis and to predict of air pollutant concentrations (PM_{10}, SO_2 and NO_2) using linear regression approach in the second largest city in the region Stara Zagora, Bulgaria - Kazanlak.

Materials and Methods

- Description of study area.

Kazanlak is a city in central Bulgaria and it is second largest city and the administrative centre of the municipality of the same name in 5th largest region in the country - Stara Zagora. The exact coordinates are 42°37′ N, 25°24′ E and altitude of 407 m.

- Air pollutants

Every material in the air which could affect human health or have a profound impact on the environment is defined as air pollutants.

Particulate matter (PM) is a complex mixture of solids and aerosols composed of small droplets of liquid, dry solid fragments, and solid cores with liquid coatings. PM_{10} (with a diameter of 10 μm or less) can enter the human respiratory system and cause acute respiratory, cardiac-associated disease and

* Corresponding author e-mail: mivanova_tru@abv.bg

mortality [6]. According to WHO [7], the annual average concentrations of PM_{10} should not exceed 15 $\mu g/m^3$, while 24-hour average exposures should not exceed 45 $\mu g/m^3$ more than 3-4 days per year.

Sulphur dioxide (SO_2) is a colourless, highly reactive gas, which is considered as an important air pollutant. The major human health concerns associated with exposure to high concentrations of SO_2 include respiratory irritation and dysfunction, aggravation of existing cardiovascular disease, damages to the eyes, mucous membranes and redness, and blisters of the skin [8]. The recommended average 24-hour emission level for SO_2 by WHO [7] is 40 $\mu g/m^3$ more than 3-4 days per year.

Nitrogen dioxide (NO_2) is a reddish-brown, highly reactive gas that is generated by oxidation of nitrogen monoxide in the atmosphere. NO_2 may cause respiratory infections, asthma, lung cancer, it can decrease the lungs' defences against bacteria making them more susceptible to infections [9]. According to WHO [7], the annual average concentrations of NO_2 should not exceed 10 $\mu g/m^3$, while 24-hour average exposures should not exceed 25 $\mu g/m^3$ more than 3-4 days per year.

- Air quality data collection

The indicators for the air pollutants PM_{10}, SO_2 and NO_2 in Kazanlak, Bulgaria from January 1, 2023 to April 30, 2024 were determined with manual sampling. Samples were taken once 24-hour intervals at the same time. The point of sampling (42°37'30" N, 25°23'43" E, altitude 381 m) was located 690 m from the central part of Kazanlak. The main sources of emission around the point were combustion processes in the industry; commercial, administrative and residential sectors; and road transport.

- Methods of analysis

PM_{10} fraction was determined directly by weight (gravimetric) method – BSS 17.2.4.20; BSS EN 12341. SO_2 was determined directly by a spectrophotometric method – BSS 17.2.4.17-83. NO_2 was determined directly by a spectrophotometric method – BSS 17.2.4.22-83.

- Time series analysis

Due to assumption that the sample data are normally distributed and have the same characteristics as the population, the descriptive statistics for the air pollutants in Kazanlak were considered.

The Shapiro-Wilk test [10] was used to check whether the data is normally distributed (H0 hypothesis) or not (H1 hypothesis). The Shapiro-Wilk test exhibiting high power, leading to good results even with a small number of observations.

The Pearson correlation coefficient [11] was used to measure the correlation between considered three air pollutants. The degree of the correlation can be identifying by the calculated Pearson correlation (r) value, where $|r| \leq 0.3$ indicated a negligible correlation; $0.3 < |r| \leq 0.5$ was a weak correlation; $0.5 < |r| \leq 0.7$ represented a moderate correlation; $0.7 < |r| \leq 0.9$ denoted a strong correlation; $0.9 < |r| \leq 1$ was fully correlated.

- Model specification

The linear regression is a statistical model which estimates the linear relationship between a dependent variable and one or more independent variables. The case of one explanatory variable is called simple linear regression [12] whose and the basic equation is: response = constant + parameter × predictor.

The *F*-test for linear regression use to test whether any of the independent variables in a linear regression model are significant. Relatively high values of *F* are strong arguments against H0 (there is no predictive relationship between considered variables) and the smaller the *p*-value the lower chance that H0 fair. For *p*-value < 0.05 to reject H0 the following condition is necessary: $F_{calculated} > F_{tabular}$.

- Statistical software.

For all statistical computing, test and graphics the software environment R version 4.3.1 was used. The statistical analysis was implemented using the ggplot2 and corrplot packages of R.

Results and Discussion

Table 1 presents the results of the descriptive statistics of the air pollutants (PM_{10}, SO_2 and NO_2) in Kazanlak from January 01, 2023 to April 30, 2024. The PM_{10} has the highest mean value of 19.09, while that of NO_2 and SO_2 have the values of 6.87 and 4.78, respectively. Hence, PM_{10} pollutant exerts the greatest influence on the status of air quality in Kazanlak. It is observed that the maximum values PM_{10} (48.94 μg/m^3) and NO_2 (34.16 μg/m^3) pollutants surpasses the recommended average 24-hour emission level WHO [7]. For the period under study there are only two values for PM_{10} and seven values for NO_2, that exceed the recommended average 24-hour emission levels for PM_{10} and NO_2. Let's note that for the entire year 2023, the days with values exceeding the recommended average 24-hour emission levels for NO_2 are only three, which does not violate WHO recommendations. The standard deviation of PM_{10} is the highest among the other two variables, thus implying that PM_{10} presents more volatile behavior than the other pollutants. SO_2 shows the lowest standard deviation and thus exhibits the most stable dynamic behavior. The skewness for the air pollutants data in Kazanlak is positive, with the greatest value at 1.97 for the SO_2 and the lowest value of 0.56 is for the PM_{10}, meaning that the data are skewed to the right. In assessing the kurtosis of air pollutants dataset in Kazanlak, SO_2 concentrations was found to have the highest value of 6.88, which indicates that its appear most frequently in unhealthy pollution events. While that of PM_{10} had the lowest value of –0.18. All the variables have a positive number for the Shapiro-Wilk statistic with a *p*-value less than 0.05. Therefore, we reject the null hypothesis that the data are normally distributed, much like [13]. Nevertheless, we can conclude that residuals of all variables are approximately normally distributed, as the skewness is between −2 and +2 and kurtosis is between −7 and +7 [14].

Table 1. Descriptive statistics results of the variables between January, 01 2023 and April, 30 2024.

Figure 1 is a graphical display of a correlation matrix between PM_{10}, SO_2 and NO_2 air pollutants. Here the positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients. A moderate positive correlation is between SO_2 and NO_2 ($r = 0.63$, $p < 0.0001$). Another group represented by PM_{10} and SO_2 ($r = 0.40$, $p < 0.0001$) and PM_{10} and NO_2 ($r = 0.33$, $p < 0.0001$) displayed a significantly weak positive correlation. Positive correlation indicates that there is a direct linear relationship between the observed variables. Some of the correlation numbers are fainter color since they are near to 0.

Figure 1. Graphical display of a correlation matrix between all considered air pollutants.

The regression analysis of the constructed model between NO_2 and SO_2 highlighted the fact that the relation between dependent and independent factor was moderate (Table 2): roughly 60% of the variance found in the response variable (NO_2) can be explained by predictor variable (SO_2). The calculated values of the *F*-test of this regression model indicated the relevance of the model ($F = 319.2$, *p*-value $< 2.2\times10^{-16}$) which had a high *F*-value and low *p*-value and standard error. Weak interdependences we were obtained in the cases of the other two regression models (between SO_2 and PM_{10}; and between NO_2 and PM_{10}): roughly 40% of the variance found in the response variable (SO_2) can be explained by predictor variable (PM_{10}) and roughly 30% of the variance found in NO_2 can be explained by PM_{10}. The regression model between NO_2 and PM_{10} had the lowest *F*-value and the greatest dispersion of data. The tabular values given by *F*-test for the three regression models were 3.94 for a probability of 0.05, which means that the resulting equations were: Fcalculated > Ftabular, consequently the H0 was rejected and the variances included in the study differed significantly between them.

Table 2. Summary statistics from linear regression models for the air pollutants.

Figure 2a-c contains graphs representing predicted data for NO_2 for each particular measurement of SO_2, for SO_2 for each particular measurement of PM_{10} and for NO_2 for each particular measurement of PM_{10} as a lower, fit, and upper predicted value, as well as a graph presenting NO_2 and SO_2 measured values.

Figure 2. Graphs of comparison between the real measurement data and the computed prediction data, based on linear regression models.

Conclusion

Based on the obtained linear regression models, the following interpretations for NO_2 and SO_2 air pollutants concentrations in Kazanlak, Bulgaria could be made: if SO_2 increase by 1%, the effect of this increase would result in an increase in NO_2 by 0.90%; if PM_{10} increase by 1%, the effect of this increase would result in an increase in SO_2 by 0.19% and if PM_{10} increase by 1%, the effect of this increase would result in an increase in NO_2 by 0.23%.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgements

This work was partially financially supported by the Scientific Projects AF_07/23 and VMF_01/24, funded by the Agricultural Faculty and Faculty of Veterinary Medicine, Trakia University. The results were presented at the XXXIV International Online Scientific Conference, June 6-7 2024, Star Zagora, Bulgaria.

References

1. Crippa M, Janssens-Maenhout G, Dentener F, Guizzardi D, Sindelarova K, Muntean M, Van Dingenen R, Granier C. Forty years of improvements in European air quality: regional policy-industry interactions with global impacts. Atmos Chem Phys 2016;16:3825–3841.
2. Reuters. No Country Met WHO Air Quality Standards in 2021, Survey Shows. URL: https://www.reuters.com/business/environment/no-country-met-who-air-quality-standards-2021-data-2022-03-22.
3. Rajat RR, Vaibhav D, Ridam G, Rahul P, Pratik G, Mukul S, Ritik J, Preetee K. Prediction of air quality index using supervised machine learning. Int J Res Appl Sci Eng Tech 2022;10:1371-1382.
4. Setiawan I. Time series air quality forecasting with R Language and R Studio. J Phys Conf Ser 2020; 1450:012064.
5. Stoyanov N, Pandelova A, Georgiev T, Kalapchiiska J, Dzhudzhev B. Forecasting of air pollution with time series and multiple regression models in Sofia, Bulgaria. J Environ Eng Landsc Manag 2023;31(3):176–185.
6. Sadeghi M, Ahmadi A, Baradaran A, Masoudipoor N, Frouzandeh S. Modeling of the relationship between the environmental air pollution, clinical risk factors, and hospital mortality due to myocardial infarction in Isfahan, Iran. J Res Med Sci 2015;20:757–62.
7. WHO (World Health Organization): Air Quality Guidelines, 2021
8. Khalaf EM, Mohammadi MJ, Sulistiyani S, Coronel AAR, Kiani F, Jalil AT, Almulla AF, Asban P, Farhadi M, Derikondi M. Effects of sulfur dioxide inhalation on human health: a review. Rev Environ Health 2022;39(2):331-337.

9. Zeng W, Zhao H, Liu R, Yan W, Qiu Y, Yang F, Shu C, Zhan Y. Association between NO_2 cumulative exposure and influenza prevalence in mountainous regions: a case study from southwest China. Environ Res 2020;189:1-9.
10. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965;52:591-611.
11. Awang NR, Elbayoumi M, Ramli NA, Yahaya AS. Diurnal variations of ground-level ozone in three port cities in Malaysia. Air Qual Atmos Health 2016;9:25–39.
12. Freedman DA (Ed). Statistical Models: theory and practice, Cambridge University Press, 2009.
13. Masseran N, Hussain SI. Copula modelling on the dynamic dependence structure of multiple air pollutant variables. Mathematics 2020;8:1-15.
14. Byrne BM (Ed). Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming, Routledge, New York, NY, USA, 2016.

Table 1. Descriptive statistics results of the variables between January, 01 2023 and April, 30 2024.

Pollutant	Mean	St. Dev.	Min	Max	Skewness	Kurtosis	SW (*p*-value)
PM_{10}	19.09	8.75	2.30	48.94	0.56	-0.18	0.97 (3.33×10^{-08})
SO_2	4.78	4.17	0.65	33.03	1.97	6.88	0.62 ($< 2.2\times10^{-16}$)
NO_2	6.87	5.96	1.00	34.16	1.73	2.48	0.75 ($< 2.2\times10^{-16}$)

Table 2. Summary statistics from linear regression models for the air pollutants.

Coefficient	Estimate	Std. error	*t*-value	*p*-value	Multiple R^2	Adj. R^2	Residual std. error	*F*-stat. (*p*-value)
$NO_2 = a_0 + a_1SO_2$								
a_0	2.559	0.320	8.001	9.11×10^{-15}**	0.597	0.596	4.629	319.2 ($< 2.2\times10^{-16}$)
a_1	0.901	0.050	7.866	$< 2\times10^{-16}$**				
$SO_2 = \beta_0 + \beta_1PM_{10}$								
β_0	1.119	0.416	2.689	0.007*	0.362	0.360	3.821	93.69 ($< 2.2\times10^{-16}$)
β_1	0.192	0.020	9.679	$< 2\times10^{-16}$**				
$NO_2 = \gamma_0 + \gamma_1PM_{10}$								
γ_0	2.556	0.613	4.170	$3.60\times^{10-05}$**	0.310	0.308	5.626	59.8 ($< 6.15\times10^{-14}$)
γ_1	0.226	0.029	7.733	6.15×10^{-14}**				

Significance codes: ** when $p \leq 0.001$; * when $p \leq 0.01$

Figure 1. Graphical display of a correlation matrix between all considered air pollutants.

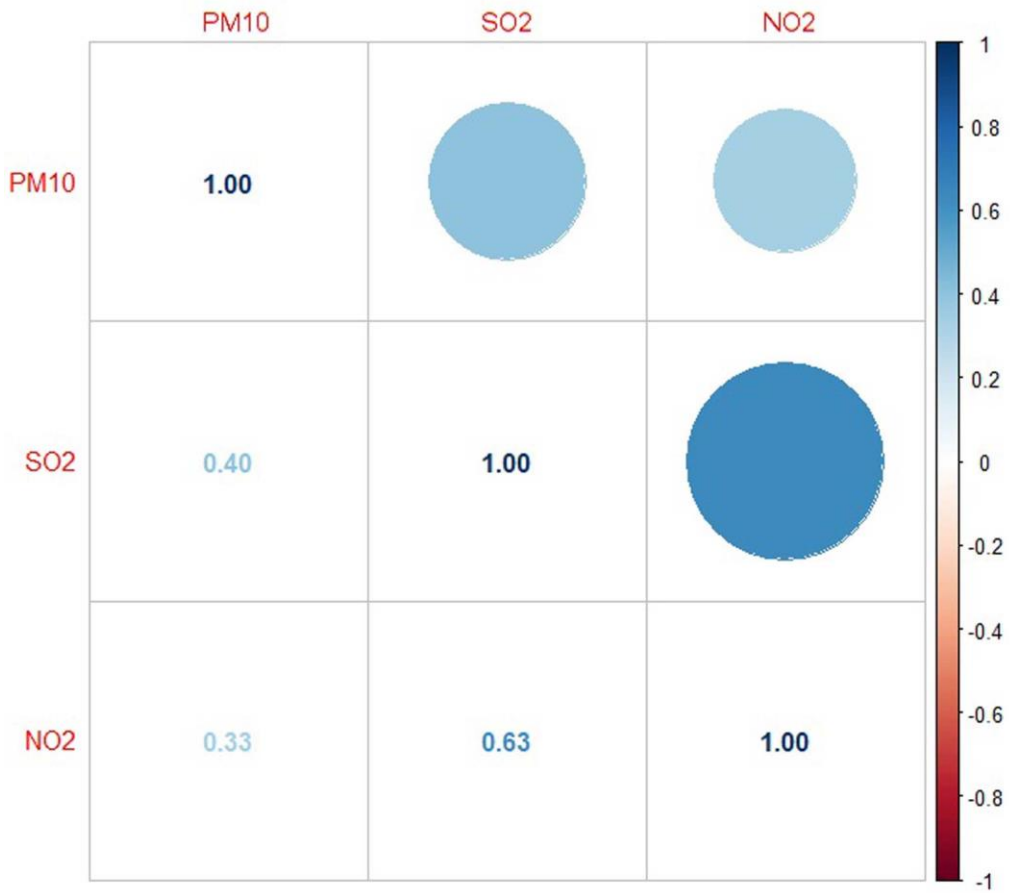

Figure 2. Graphs of comparison between the real measurement data and the computed prediction data, based on linear regression models.