Multivariate Analysis for Modeling of Air Pollutants and Ozone Concentration in Dimitrovgrad, Bulgaria

Snezhana Georgieva Gocheva-Ilieva, Atanas Valev Ivanov, and Iliycho Petkov Iliev

***Abstract*—Air pollution is one of the key problems in urban areas and its investigation is vital both for people's health but also for the environment as a whole. In particular, ozone is a secondary air pollutant with concentrations dependent mainly on changes in the levels of other pollutants and meteorological conditions within a given region. This paper presents a multivariate analysis of hourly data on 9 air pollutants and 6 meteorological variables in the town of Dimitrovgrad, Bulgaria over a period of 7 years and 3 months. Yeo-Johnson power transformation is applied to each air pollutant variable to improve normality and symmetry. The dominant patterns in the considered data are examined with the help of Principal Component Analysis (PCA) and factor analysis. Furthermore, particular focus is given to determining the concentration levels of ozone in relation to the other air pollutants and 6 meteorological parameters using multiple linear regression. The fitting of the obtained models with coefficients of determination R2 over 78% are obtained. The results are interpreted.**

***Index Terms*—Air pollution modeling, factor analysis, multiple linear regression, ozone concentration, principal component analysis (PCA).**

I. Introduction

The monitoring, investigation and control of ambient air quality is a topical issue, very important for preserving the human health and the environment. Directives and regulatory restrictions are established in all European countries and worldwide for permissible concentrations of air pollutants [1]-[3]. 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 availability of a huge amount of collected data allows their statistical examination and makes it possible to find significant patterns, as well as dependencies within the data enabling the prediction of future states.

In literature, numerous similar investigations have been carried out in recent years, applying various mathematical methods supported by different statistical software. Multivariate statistical analysis is a well-established approach in this field. Recent studies where Principal Component Analysis (PCA), factor analysis, regression analysis and other methods were applied include for example [4]–[8]. Other popular techniques are the stochastic Box-Jenkins ARIMA methods [9], neural networks [10], etc.

The objective of this paper is to investigate the relationships between large number of air pollutants in Dimitrovgrad, a town in Bulgaria, over an extended period of 7 years and 3 months. The specific goals of the study are: 1) Establishing the existence and determining the type of correlations between the investigated air pollutants; 2) Defining patterns and possible groups of pollutants which are closely related, and classifying the pollutants; 3) Obtaining multivariate regression models which describe the changes in ozone pollution levels in relation to the other pollutants and meteorological data.

Manuscript received June 2, 2015; revised December 29, 2015. This work was supported in part by Plovdiv University "Paisii Hilendarski" NPD under Grant NI15-FMI-004.

S. G. Gocheva-Ilieva and A. V. Ivanov are with the Department of Applied Mathematics and Modeling, Plovdiv University "Paisii Hilendarski", 24 Tsar Asen Street, 4000 Plovdiv, Bulgaria (e-mails: snow@uni-plovdiv.bg, aivanov@uni-plovdiv.bg).

I. P. Iliev is with the Physics Department, Technical University-Sofia, Branch Plovdiv, 25 Tsanko Diustabanov Street, 4000 Plovdiv, Bulgaria (e-mail: iliev55@abv.bg).

II. Materials and Methods

A. Study Area and Data Description

We examine air quality in the town of Dimitrovgrad, a typical urban region in South-central Bulgaria, 220 km away from the capital city of Sofia. The town is located in the Thracian valley on the banks of the Maritza river at an altitude of 125 m above sea level. It has about 40 000 inhabitants.

The study was carried out based on hourly data about concentration of air pollutants between 1st January 2007 and 7 March 2014. Measurements were taken by an automated monitoring station in the town run by the official Executive Environment Agency.

The following 9 air pollutants are considered: nitrogen oxides (NOx, ppb); nitrogen dioxide (NO_2, $\mu g/m^3$), nitrogen oxide (NO, $\mu g/m^3$), ozone (O_3, $\mu g/m^3$), carbon monoxide (CO, $\mu g/m^3$), sulphur dioxide (SO_2, $\mu g/m^3$), hydrogen sulfide (H2S, $\mu g/m^3$), ammonia, or azane, a compound of nitrogen and hydrogen (NH_3, $\mu g/m^3$), particulate matter with diameter of 10 micrometres or less (PM10, $\mu g/m^3$).

Basic descriptive statistics of the data are given in Table I, columns 1-8.

The following 6 meteorological variables are also used: wind speed (WS, m/s), wind direction (SIGMA, degree), air humidity (HUMIDITY, %), air temperature (TEMP, ^{0}C), sun radiation (GSR, W/m^2), atmospheric pressure (PRESSURE, mbar).

We have to note that although the mean value of ozone is not very high, the examined data indicate some values exceeding systematically the permissible limits established by the health regulations as set out in [1]-[3]. This is the reason to model this particular secondary air pollutant.

TABLE I: DESCRIPTIVE STATISTICS OF THE OBSERVED AND TRANSFORMED AIR POLLUTANTS

Variable	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis	Transformed variable	λ	Skewness of transf. variable	Kurtosis of transf. variable	JB test
O3	60236	0.00	1094	49.75	35.353	3.674	77.472	[O3]	0.44	-0.011	0.049	8
NOx	61612	0.00	468	14.63	23.318	5.676	48.422	[NOx]	-1.5	0.187	-0.963	2801
NO	60621	0.00	451	6.47	21.323	7.541	77.478	[NO]	-2	1.400	0.913	22743
NO2	61013	0.00	212	18.42	16.739	2.248	7.635	[NO2]	-0.15	0.167	-0.710	1614
CO	58110	0.00	10	0.53	0.719	3.665	20.947	[CO]	-0.65	0.245	-0.828	2432
SO2	61641	0.00	1662	31.40	60.123	6.388	73.867	[SO2]	-2	0.374	-1.092	4597
H2S	60570	0.00	0.06	0.003	0.003	5.141	54.114	[H2S]	-0.15	0.021	0.025	6
NH3	60570	0.00	0.16	0.003	0.004	6.724	144.895	[NH3]	-0.4	0.391	-0.670	2781
PM10	61261	0.03	896	57.08	57.165	3.738	22.080	[PM10]	-0.6	0.057	-0.514	726

Std. error of Skewness for all variables is 0.010; Std. error of Kurtosis for all variables is 0.020. Kurtosis is the excess kurtosis.

B. Data Transformation

In principle, applying multivariate analysis requires an assumption for normal or close to normal distribution of participating variables, as well as other assumptions [11]. As shown in Table I, the Skewness and the Kurtosis of the considered data are significantly different from zero, therefore, the normality condition is strongly violated. In such cases, it is advisable to improve the distribution by a suitable preliminary data transformation [9], [12]. Firstly, the variables of all air pollutants were standardized. Then, the power transformation of Yeo-Johnson [13] was applied by the formula:

$$\psi_{YJ}(\lambda, x) = \begin{cases} \dfrac{(x+1)^{\lambda}-1}{\lambda} & x \geq 0,\ \lambda \neq 0 \\ \log(x+1) & x \geq 0,\ \lambda = 0 \\ -\dfrac{(-x+1)^{2-\lambda}-1}{2-\lambda} & x < 0,\ \lambda \neq 2 \\ -\log(-x+1) & x < 0,\ \lambda = 2 \end{cases} \quad (1)$$

where $\lambda \in [-2, 2]$ is a parameter. The optimal values of λ were chosen to give the smallest possible value of the Jarque-Bera test of normality [14], defined as

$$JB = n\left[\frac{Sk^2}{6} + \frac{Ku^2}{24}\right], \quad (2)$$

where *n* is the number of cases, *Sk* and *Ku* are the Skewness and the excess Kurtosis of the sample, respectively.

The last three columns of Table I show the obtained optimal values λ and Skewness and Kurtosis of the transformed variables.

C. Multivariate Statistical Methods

The examination of the correlation matrices for the initial and the transformed data of the 9 pollutants indicates the presence of high multicollinearity. To resolve this problem and to classify the data, we use the well-known PCA method [15]. PCA allows the extraction of linearly independent principal components (PCs), equal in number to the number of output variables. Factor analysis is also used to identify patterns in data using the main extracted PCs, which group together the variables and account for the greater part of the total sample variance.

In order to model ozone concentrations, the multiple linear regression (MLR) method is applied resulting in an explicit dependence of the type.

$$[O3] = b_0 + b_1 X_1 + ... + b_m X_m \ , \quad (3)$$

where $[O3]$ is the transformed ozone variable, $b_0, b_1, ..., b_m$ are regression coefficients and $X_1, X_2, ..., X_m$ are the predictors (independent variables). In (3) the PCs obtained from the transformed variables for pollutants and/or meteorological variables could be used as predictors. This type of dependence corresponds to the chemical processes which lead to ozone formation in urban areas, taking into account the influence of ozone precursors and meteorological conditions [1].

Missing values are treated listwise.

Calculations are performed using the IBM SPSS statistical software.

III. Results from Principal Component Analysis for Air Pollutant Variables

Further, we will work with both initial and transformed variables to compare results. Within the first step of PCA the correlation matrices were calculated as presented in Table II and Table III.

All columns contain coefficients over 0.3, with the largest correlation coefficient being that between [NOx] and [NO2], equal to 0.957. All correlation coefficients of [O3] (ozone) with other variables are negative, which corresponds to the nature of chemical reactions since ozone is formed from other pollutants, i.e. its concentration is inversely proportional to their own. As expected, the highest negative correlations are those with nitrogen oxides, nitrogen dioxide and nitrogen oxide.

The relatively high absolute values of correlation coefficients in Tables II and III, and the small values of the determinants indicate the presence of high multicollinearity. This result is weaker in the second case.

An adequacy test was also performed for factor analysis, with the KMO test yielding a value of KMO=0.815 > 0.5 and Bartlett's test significance, equal to 0.000. This shows that the factor analysis is adequate.

The next step of the PCA method is to generate PCs resulting from the 9 transformed variables. Table IV shows

the calculated eigenvalues and the distribution of total variance. From Table IV it can be observed that the last eigenvalue is very small and could be ignored [11], [15]. With the presence of multicollinearity the number of PCs is less than 8.

TABLE II: PEARSON CORRELATIONS OF THE INITIAL AIR POLLUTNTS

Variable	O3	NOx	NO	NO2	CO	SO2	H2S	NH3	PM10
O3	1	-0.418	-0.317	-0.508	-0.348	0.011	-0.267	-0.163	-0.326
NOx		1	0.957	0.813	0.755	0.260	0.572	0.437	0.656
NO			1	0.611	0.731	0.208	0.553	0.374	0.586
NO2				1	0.600	0.293	0.454	0.447	0.620
CO					1	0.304	0.587	0.323	0.672
SO2						1	0.329	0.098	0.336
H2S							1	0.255	0.545
NH3								1	0.373
PM10									1

Significance (1-tailed) for all correlation coefficients is 0.000 in exception of (SO2, O3), which maximums are up to 0.007. Determinant = 5.56E-06.

TABLE III: PEARSON CORRELATION TABLE OF THE TRANSFORMED VARIABLES OF AIR POLLUTANTS

Variable	[O3]	[NOx]	[NO]	[NO2]	[CO]	[SO2}	[H2S]	{NH3]	[PM10]
[O3]	1	-0.700	-0.644	-0.635	-0.399	-0.096	-0.308	-0.247	-0.411
[NOx]		1	0.756	0.949	0.450	0.348	0.336	0.367	0.595
[NO]			1	0.640	0.502	0.246	0.357	0.340	0.522
[NO2]				1	0.413	0.370	0.309	0.344	0.573
[CO]					1	0.332	0.297	0.215	0.424
[SO2]						1	0.242	0.086	0.420
[H2S]							1	0.258	0.337
[NH3]								1	0.337
[PM10]									1

Significance (1-tailed) for all correlation coefficients is 0.000. Determinant = 0.004.

TABLE IV: TOTAL VARIANCE EXPLAINED.

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
PC1	4.492	49.907	49.907
PC2	1.043	11.586	61.493
PC3	0.891	9.897	71.390
PC4	0.760	8.442	79.832
PC5	0.655	7.282	87.114
PC6	0.462	5.136	92.250
PC7	0.340	3.774	96.024
PC8	0.322	3.581	99.605
PC9	0.036	0.395	100.000

Extraction Method: Principal Component Analysis.

To classify the air pollutants and discover the dominant patterns in the dataset we perform factor analysis. Varimax rotation did not yield well differentiated components. For this reason we applied Promax rotation. The optimal factor solution with all 9 pollutants contains 7 factors, which account for 96.024% of total variance. The resulting rotated solution with 7 factors is given in Table V. We have to add, that all the variance inflation factors (VIF) are less than 2.85, which indicate that the obtained PCs do not correlate one with another.

Table V clearly shows that all PCs are well differentiated. PC1 groups together [NO2], [Nox], and [O3]. All other variables are individual factors.

Analogically to the previous analysis, after excluding ozone, we apply PCA and get 7 uncorrelated PCs, which are used in the subsequent analyses. These PCs account for 99.548% of the total variance. The loadings of the rotated matrix are shown in Table VI.

IV. RESULTS FROM MULTIPLE LINEAR REGRESSION ANALYSIS

The next goal is to find the explicit dependence between ozone and the other pollutants, with and without the meteorological data.

Due to the multicollinearity of the variables, the direct application of multiple linear regression to non-transformed or transformed variables is not recommended and give unsatisfactory results. This can be overcome using a well-known technique, namely the extraction of principal components which are not mutually correlated. The obtained new variables can be used to find regression models. This mixed regression approach is known as Principal Component Regression (PCR) [11].

In previous section 3, after excluding ozone, we obtained 7 uncorrelated PCs, which are used in subsequent regression analyses. In addition to these variables the six meteorological variables are also included as predictors to obtain regression equations.

Multivariate regressions are performed using the Stepwise method in SPSS. The statistical significance is established at level $\alpha = 0.05$.

The obtained standardized regression equation using the 7 extracted PCs, according the PCA (see Table 6) has the form

$$[O3] = -0.502PC1 - 0.092PC2 + 0.054PC3 + 0.202PC4 - 0.077PC5 - 0.008PC6 - 0.288PC7 \quad (4)$$

The coefficient of determination of equation (4) is $R^2 = 0.551$. All coefficients, as well as the ANOVA of the model are statistically significant.

The relative influence of ozone precursors on the examined data is defined. The results show the strongest influence is that of $PC1 = \{NO2, NOx\}$ and $PC7 = \{NO\}$, followed by $PC4 = \{SO2\}$. The remaining pollutants have weaker influence.

TABLE V: PRINCIPAL COMPONENT PATTERN MATRIX FOR 9 AIR POLLUTANTS

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
[NO2]	**1.052**	0.083	0.035	-0.009	-0.020	0.020	-0.175
[NOx]	**0.897**	0.072	0.039	-0.023	-0025	0.023	0.078
[O3]	**-0.636**	0.239	0.112	-0.053	-0.072	0.057	-0.358
[SO2]	0.020	**0.974**	-0.041	0.018	0.026	-0.020	0.077
[NH3]	0.011	-0.043	**0.985**	0.009	0.013	-0.010	0.037
[CO]	-0.001	0.020	0.009	**0.998**	-0.006	0.005	-0.013
[H2S]	-0.005	0.027	0.013	-0.005	**1.000**	0.007	-0.023
[PM10]	0.021	-0.021	-0.010	0.006	0.007	**0.985**	0.021
[NO]	0.053	0.113	0.053	-0.018	-0.032	0.029	**0.916**

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization. Rotation converged in 7 iterations.

TABLE VI: FACTOR ANALYSIS PATTERN MATRIX WITH 7 FACTORS FOR 8 AIR POLLUTANTS

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
[NO2]	**1.052**	0.005	0.000	0.009	0.001	0.001	-0.102
[NOx]	**0.890**	-0.005	0.001	-0.010	0.000	0.005	0.145
[CO]	0.003	**0.997**	0.000	0.000	0.000	0.001	0.003
[NH3]	0.001	0.000	**0.999**	0.000	0.000	0.000	0.000
[SO2]	0.002	0.000	0.000	**0.998**	0.000	0.001	0.002
[H2S]	0.001	0.000	0.000	0.000	**1.000**	0.000	0.000
[PM10]	0.014	0.001	0.001	0.002	0.000	**0.**990	0.001
[NO]	0.048	0.005	0.000	0.002	0.001	0.001	**0.963**

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization. Rotation converged in 6 iterations.

It is well-known that ozone concentration is strongly dependent on meteorological conditions which influences chemical reactions leading to its formation. The next model is derived using the six meteorological variables. The resulting standardized equation has the following form:

$$[O3] = -0.410\,HUMIDITY + 0.295\,TEMP + 0.260\,WS - 0.115\,SIGMA + 0.098\,PRESSURE + 0.068\,GSR \quad (5)$$

The corresponding coefficient of determination is $R^2 = 0.635$. It becomes clear from (5) that as a whole the main contribution in the examined interaction is that of low air humidity, air temperature and wind speed.

Finally, all 7 PCs from Table 6 and 6 meteorological predictors are used to simultaneously take into account the precursors and the meteorological data. The resulting standardized regression equation is

$$[O3] = -0.262\,PC1 - 0.020\,PC2 + 0.035\,PC3 + 0.159\,PC4 - 0.009\,PC5 + 0.006\,PC6 - 0.265\,PC7 - 0.345\,HUMIDITY + 0.161\,TEMP + 0.103\,WS - 0.082\,SIGMA + 0.066\,PRESSURE + 0.080\,GSR \quad (6)$$

The coefficient of determination of model (6) is $R^2 = 0.783$. The dominant part of the equation is due to the following principal components and predictors: $PC1 = \{NO2, NOx\}$, $PC7 = \{NO\}$, $HUMIDITY$, $TEMP$, and $PC4 = \{SO2\}$.

V. DISCUSSION AND CONCLUSION

With the help of PCA and the correlation matrix, 8 PCs were derived and classified according to their relative contribution to the total level of air pollution. Nitrogen oxides (NOx, NO2, NO) play a dominant part. This result is explained by the presence of a large nitrogen fertilizer production plant in the town of Dimitrovgrad, Bulgaria, which is the main source of industrial air pollution in the city, along with road traffic.

The obtained regression equation (6) shows that the combined contribution of ozone precursors and meteorological data account for up to 78% of the concentration of this pollutant. The other two equations (4) and (5) show lower values, but these are significant for the clear differentiation of the relative participation of each pollutant in overall ozone concentration.

It is necessary to note, that as it is well known, there are differences between daytime and nighttime ozone concentrations. This is to be examined in a future paper.

REFERENCES

[1] European Commission. (2013). *Environment. Air Quality Standards*. [Online]. Available: http://ec.europa.eu/environment/air/quality/standards.htm

[2] 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, *Official Journal of the European Union* L 152/1, 2008.

[3] European Environmental Agency. (2014). Air Quality in Europe — 2014 Report. [Online]. Available: http://www.eea.europa.eu/publications/air-quality-in-europe-2014

[4] A. S. Kaplunovsky, "Factor analysis in environmental studies," *HAIT Journal of Science and Engineering B*, vol. 2, no. 1–2, pp. 54–94, 2005.

[5] J. P. Shi and R. M. Harrison, "Regression modelling of hourly NOx and NO2 concentrations in urban air in London," *Atmospheric Environment*, vol. 31, pp. 4081-4094, 1997.

[6] A. Lengyel, K. Héberger, L. Paksy, O. Bánhidi, and R. Rajkó, "Prediction of ozone concentration in ambient air using multivariate methods," *Chemosphere*, vol. 57, pp. 889-896, 2004.

[7] V. Gvozdić, E. Kovač-Andrić, and J. Brana, "Influence of meteorological factors NO2, SO2, CO and PM10 on the concentration of O3 in the urban atmosphere of Eastern Croatia", *Environmental Modeling & Assessment*, vol. 16, no. 5, pp. 491-501, 2011.

[8] T. W. Chan, and M. Mozurkewich, "Application of absolute principal component analysis to size distribution data: identification of particle origins," *Atmospheric Chemistry and Physics*, vol. 7, pp. 887-897, 2007.

[9] G. E. P. Box, and G. M. Jenkins, *Time Series Analysis, Forecasting and Control*, San Francisco: Holden Day, 1976.

[10] A. Azid, H. Juahir, M. E. Toriman, M. K. A. Kamarudin, A. Shakir, M. Saudi, C. N. C. Hasnam, N. A. A. Aziz, F. Azaman, M. T. Latif, S. F. M. Zainuddin, M. R. Osman, and M. Yamin, "Prediction of the level of air pollution using principal component analysis and artificial neural Network techniques: A case study in Malaysia," *Water, Air, & Soil Pollution*, vol. 225, pp. 2063-2074, 2014.
[11] A. Izenman, *Modern Multivariate Statistical Techniques*, New York: Springer, 2008.
[12] S. Wold, K. Sbensen and P. Geladi, "Principal component analysis," *Chemometrics and Intelligence Laboratory Systems*, vol. 2, pp. 37-52, 1987.
[13] I. K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954-959, 2000.
[14] C. Jarque and A. Bera, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Economics Letters*, vol. 6, pp. 255-259, 1980.
[15] I. T. Jolliffe, *Principal* Component *Analysis*, 2nd ed., New York: Springer, 2002.

Snezhana G. Gocheva-Ilieva was born in 1950. She graduated in mathematics and obtained her M.Sc. degree in computational mathematics from Sofia University St. Kliment Ohridski, Sofia, Bulgaria in 1973. She completed her Ph.D. degree in physics and mathematics from Taras Shevchenko National University of Kyiv, Ukraine, in 1981.

She is currently working as a full professor of applied mathematics at the Department of Applied Mathematics and Modeling from Paisii Hilendarski University of Plovdiv, Plovdiv, Bulgaria and teaches at the graduate and master courses applied mathematics and computer science. She has published over 115 articles in international journals, 12 books and has presented many research papers at international conferences. Her research interests include various fields of mathematical modeling in physics and engineering, modeling in environmental science, applied computational statistics, predictive data mining techniques, and more.

Atanas V. Ivanov was born in 1986. He graduated in mathematics and obtained his M.Sc. degree from Paisii Hilendarski University of Plovdiv, Plovdiv, Bulgaria in 2010. He completed his Ph.D. degree in applied mathematics from the same University in 2015.

He is currently working as an assistant professor of applied mathematics at the Department of Applied Mathematics and Modeling at Paisii Hilendarski University of Plovdiv, Plovdiv, Bulgaria and teaches at the graduate courses in applied mathematics. He has published 6 articles in international journals, and has presented some research papers at international conferences. His research interests include mathematical modeling in environmental science, applied statistics, and predictive data mining techniques.

Iliycho P. Iliev was born in 1955. He graduated in electrical engineering and obtained his M.Sc. degree of engineer from Moscow Power Engineering Institute, Moscow, Russia in 1982. He completed his Ph.D. degree in physics from Bulgarian Academy of Sciences in 2002.

He is currently working as an associate professor of physics at Physics Department, Technical University Sofia, branch Plovdiv, Plovdiv, Bulgaria and teaches at the graduate courses in physics and engineering. He has published over 100 articles in international journals, 6 books and has presented many research papers at international conferences. His research interests include laser physics, modeling in physics and engineering, statistical modeling, and others.