A Complex Analysis Employing ARIMA Model and Statistical Methods on Air Pollutants Recorded in Ploiesti, Romania

ALIN POHOATA*, EMIL LUNGU
Valahia University of Targoviste, 13 Sinaia Alley, 130004, Targoviste, Romania

Air pollution is an everyday issue, very relevant to public authorities, requiring control and monitoring to provide data for decision-making policies. The objective of this study was to evaluate the air quality in Ploiesti city, Romania and to observe the advantages and limitations of the some statistical methods used in forecasting air quality. Data for six air quality parameters collected at monitoring stations in Ploiesti during the 2013 year were statistically analyzed. Principal component analysis (PCA) was used to provide a relevant description in factors that can be explained in terms of different sources of air pollution. The measured pollutants data were statistically analyzed using the auto-regressive integrated moving average (ARIMA) method in order to assess the efficiency of using this method in forecasting the environmental air quality. The results revealed that ARIMA method has some limitations and do not produce satisfactory results for certain air pollutants such as PM_{10} and CO, even the forecasted period is short. By comparison, the ARIMA model obtained for NO_x, NO_2, or O_3 time series, provides good results, with relative errors around 5%.

Keywords: air pollution, principal component analysis, forecasting, ARIMA

Monitoring and control of air pollution became an important issue for public authorities, being required for air quality planning and related decision-making policies.

Air pollution in urban areas is commonly characterized using several pollutants. Particulate matter (PM), ground-level ozone (O_3), sulfur dioxide (SO_2), nitric oxide (NO), nitrogen dioxide (NO_2) or carbon monoxide (CO) are some of the most commonly monitored pollutants, having in view that they can have a serious impact on human health and on the environment. Consequently, the development of rigorous mathematical models able to run with the versatile behavior of air pollutants in order to predict the air quality is an important area of research.

Statistical analysis is an important stage of transition from raw data obtained by monitoring activity, towards a system of synthetic indicators according to the behavior of the studied phenomenon, including techniques by which data are transformed into explanatory information.

Data processing techniques consist of air quality temporal data analysis (time series), highlighting extremes, frequency distribution curves and probability of peaks. Time-series modeling is an instrument that was successfully used in the analysis of environmental pollutants concentration sequences [2-4]. Changing behavior of air pollution through time was studied by many authors using different techniques [5,6]. There are authors that have tried to relate air pollution to human health using various techniques of time series analysis [7,8].

In the last years, many governmental measures were implemented for air quality control in many regions of Romania to cope with the European Union regulations. Despite of these efforts air quality, especially in major urban areas, are not satisfactory. Ploiesti is one of Romanian main urban areas with a high industrial development, especially with intensive oil processing and refining industry, oil extraction equipment and machinery, manufacturing of rubber and plastic products, detergents etc. Consequently, there are important stationary emission sources that contribute to the global emissions of air pollutants in the residential areas of the urban agglomeration. Ploiesti city has an increased car traffic and therefore, mobile sources have also a significant contribution to the poor air quality [6]. In this context, the management of air quality in Ploiesti urban area became a sensitive issue for the public authorities.

In this paper, several statistical methods for time series analysis of air pollution using daily averages concentrations are presented. The objective of this study was to evaluate the air quality in a major urban-industrial area from Romania, i.e., Ploiesti city. Particulate matter (PM), nitrogen oxides (NO_2, NO_x), sulfur dioxide (SO_2), carbon monoxide (CO) and ground-level ozone (O_3) emissions, recorded in 2013 in Ploiesti city are used as for the analysis. The measured pollutants data were statistically analyzed using the auto-regressive integrated moving average (ARIMA) method. The results were focused on the efficiency of using this method in forecasting the environmental air quality.

Experimental part

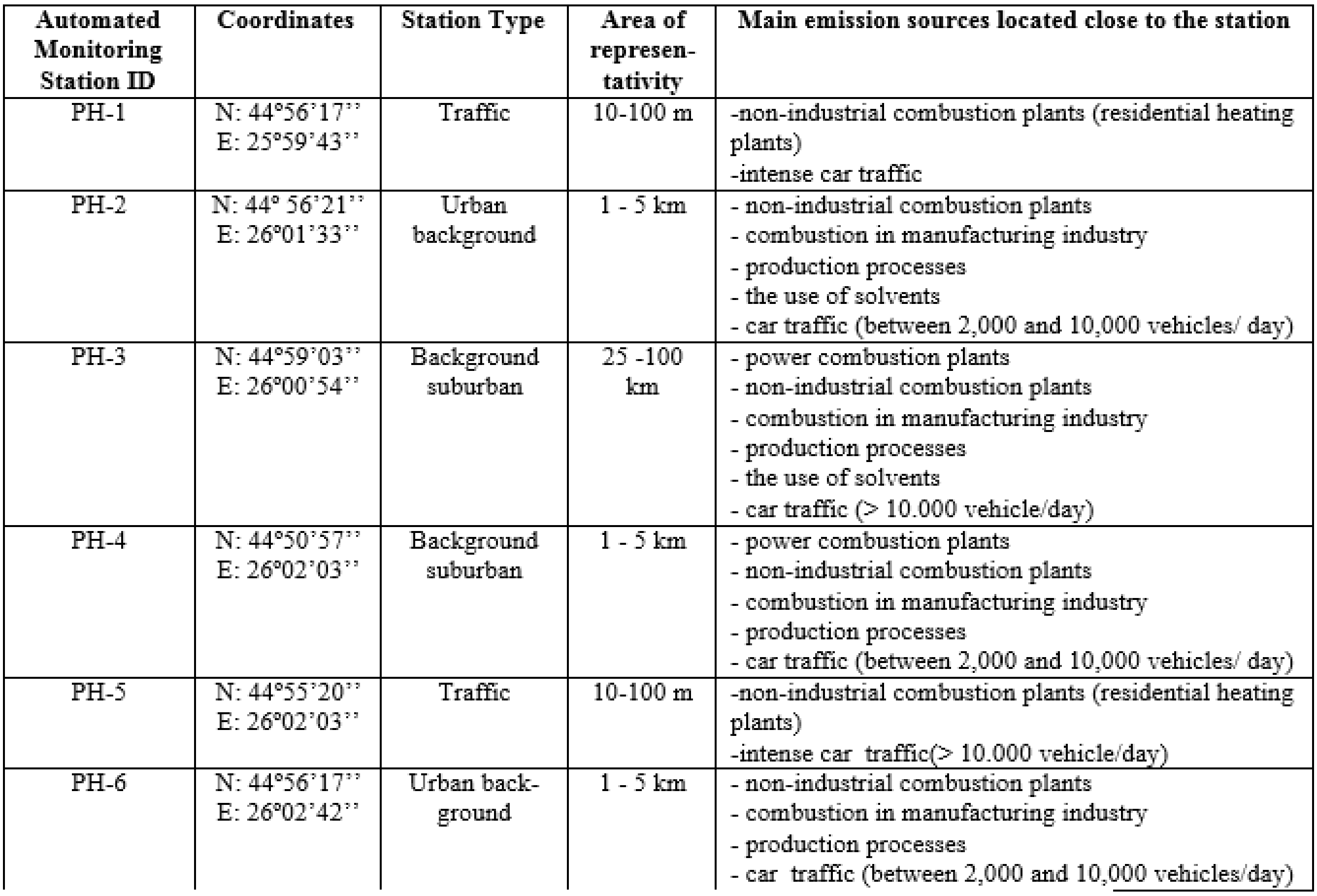
The study area is Ploiesti city, located in Southeast of Romania (44°562' 243" N Lat.; 26°012'003" E Long.). Currently, there are six automated air quality monitoring stations in operation in Ploiesti, which are part of National Network for Air Quality Monitoring (RNMCA): two traffic stations (PH-1 and PH-5), two urban background stations (PH-2 and PH-6), one suburban background station (PH-3) and one industrial station (PH-4).

The stations are equipped with automatic analyzers that continuously measure the ambient concentrations of the main atmospheric pollutants: sulfur dioxide (SO), nitrogen oxides (NO_2, NO_x), carbon monoxide (CO), ozone (O_3) particulate matter (PM_{10} and $PM_{2.5}$).

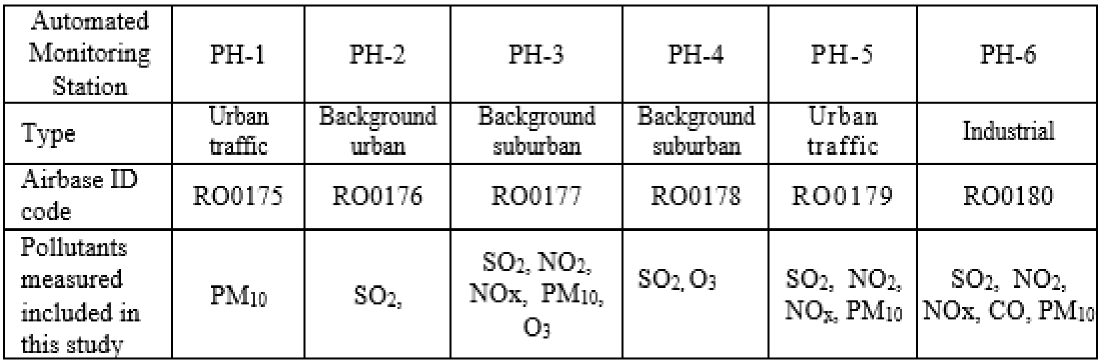
In this study, the recorded raw values of pollutant concentrations were extracted from AirBase (http://acm.eionet.europa.eu/databases/airbase/), the European air quality database maintained by the European Environmental Agency, which consists of multi-annual time series of air quality measurement data and statistics for a number of air pollutants, submitted by participating countries throughout Europe. The raw data resulted from the air quality surveillance performed by the automatic

* email: alinpohoata@yahoo.com

Table 1
CHARACTERISTICS OF THE AIR QUALITY MONITORING STATIONS FROM PLOIESTI [9]



Automated Monitoring Station ID	Coordinates	Station Type	Area of represen-tativity	Main emission sources located close to the station
PH-1	N: 44°56'17'' E: 25°59'43''	Traffic	10-100 m	-non-industrial combustion plants (residential heating plants) -intense car traffic
PH-2	N: 44° 56'21'' E: 26°01'33''	Urban background	1 - 5 km	- non-industrial combustion plants - combustion in manufacturing industry - production processes - the use of solvents - car traffic (between 2,000 and 10,000 vehicles/ day)
PH-3	N: 44°59'03'' E: 26°00'54''	Background suburban	25 -100 km	- power combustion plants - non-industrial combustion plants - combustion in manufacturing industry - production processes - the use of solvents - car traffic (> 10.000 vehicle/day)
PH-4	N: 44°50'57'' E: 26°02'03''	Background suburban	1 - 5 km	- power combustion plants - non-industrial combustion plants - combustion in manufacturing industry - production processes - car traffic (between 2,000 and 10,000 vehicles/ day)
PH-5	N: 44°55'20'' E: 26°02'03''	Traffic	10-100 m	-non-industrial combustion plants (residential heating plants) -intense car traffic(> 10.000 vehicle/day)
PH-6	N: 44°56'17'' E: 26°02'42''	Urban back-ground	1 - 5 km	- non-industrial combustion plants - combustion in manufacturing industry - production processes - car traffic (between 2,000 and 10,000 vehicles/day)



Automated Monitoring Station	PH-1	PH-2	PH-3	PH-4	PH-5	PH-6
Type	Urban traffic	Background urban	Background suburban	Background suburban	Urban traffic	Industrial
Airbase ID code	RO0175	RO0176	RO0177	RO0178	RO0179	RO0180
Pollutants measured included in this study	PM_{10}	SO_2,	SO_2, NO_2, NOx, PM_{10}, O_3	SO_2, O_3	SO_2, NO_2, NO_x, PM_{10}	SO_2, NO_2, NOx, CO, PM_{10}

Table 2
CORRESPONDENCE BETWEEN THE AIR POLLUTANT AND THE MONITORING STATION USED TO EXTRACT RAW DATA

monitoring stations from Ploiesti during the 2013 year (from January, 1st to December, 31st) were statistically analyzed. The air pollutants included in the present work were sulfur dioxide (SO_2), nitrogen oxides (NO_2, NO_x), carbon monoxide (CO), ozone (O_3) and particulate matter (PM_{10}). The selection of the air pollutants was made considering the availability of data existing in the AirBase recorded by the monitoring stations in Ploiesti. Table 2 shows the correspondence between the pollutant and the monitoring station where the analyzed time series were extracted. The status, temporal and spatial variability, and correlation of monitored parameters were established.

We computed the daily average of the recorded values and we analyzed the corresponding time series.

Air quality time series consist of complex linear and non-linear patterns and are difficult to be forecasted [10]. Air quality forecasting is based on statistical approaches or various mathematical models. Box–Jenkins Time Series (ARIMA) models have been often applied to air quality forecasting in urban areas [10].

ARMA models are a combination of auto-regressive models **AR** and mobile average **MA** models. Box and Jenkins introduced this type of model in 1970 [11]. ARMA model of order (*p, q*) is denoted by ARMA(*p,q*), and is described by the following relation:

$$Xt = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + e_t + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q}$$

where $\{\alpha_i\}$ are AR parameters, $\{\beta_j\}$ are MA parameters, p is the order of AR process and q is the order of MA process. X_t are the time series terms and e_t represents error terms. The error terms are white noise, meaning they are independent variables, identically distributed with normal distribution and zero mean. The above relation can be written as:

$$\left(1 - \sum_{i=1}^{p} \alpha_i L^i\right) X_t = \left(1 - \sum_{j=1}^{q} \beta_j L^j\right) e_t$$

where L^i is the lag operator $L^i X_t = X_{t-i}$.

A more complex development of ARMA(*p,q*) which takes into consideration the seasonal variations is **ARIMA.** The ARIMA(*p,d,q*) model can be expressed as:

$$\left(1-\sum_{i=1}^{p}\alpha_i L^i\right)\left(1-L^i\right)^d X_t = \left(1-\sum_{j=1}^{q}\beta_j L^j\right)e_t$$

where d is a positive integer that controls the differentiation level.

As any prediction method, ARIMA has its own limitations. The model is suited only for time series with variance, mean and autocorrelation approximately constant through time. The number of input observations data should be at least 50. In addition, the values of the estimated parameters are assumed to be constant throughout the series.

ARMA and ARIMA models were widely used in the prediction of the airborne pollution [12,13].

Results and discussions

Statistical Analysis

The analyzed time series consist of daily average of sulfur dioxide (SO_2), nitrogen oxides (NO_2, NO_x), carbon monoxide (CO), ozone (O_3) and particulate matter (PM_{10}) and were obtained using preprocessing of raw data recorded by the automatic analyzers from monitoring stations. Data were statistically analyzed in order to assess the evolution (the status and variability) of pollutants' concentrations. Figure 1 presents the variation of concentrations of air pollutants of concern in the year 2013.

In descriptive statistics, a box plot is a quick way to examine a numerical set of data by a graphically representation, using their quartiles. Figure 2 presents the Tukey box plots of daily average of concentrations for the analyzed pollutants (SO_2, PM_{10}, O_3, NO_x, NO_2, CO) recorded in Ploiesti city during the 2013 year. The boxes present the median, the first and third quartiles, while the whiskers represent the minimum and maximum values. The lowest and highest occurring values within this limit are drawn as bars of the *whiskers*, and the outliers as individual points. The lowest datum lies within 1.5 *inter quartile range* (IQR) of the lower quartile, and the highest datum lies within 1.5 IQR of the upper quartile. All the data which are not included between the whiskers are plotted as outliers and are defined as observations that fall below Q1 - 1.5(IQR) or above Q3 + 1.5(IQR).

The summary statistics are resumed in table 3. As it can be noticed from skewness and kurtosis values, the distribution is non-normal for the analyzed pollutants. In addition, the high value of variation coefficient (CV%), together with the maximum and minimum values, depicts a very spread distribution for CO.

Before applying a correlation test, data distributions for each air pollutant were checked to establish if the data are normally distributed. Distributions of data were analyzed using simple graphical methods and statistics tests. The graphical methods used in this study include histograms and scatter diagrams. In addition to graphical presentation, the Shapiro-Wilk test has also been applied to estimate the air pollutants normality.

It was found that all pollutants distributions were non-normal (Shapiro-Wilk test $p < 0.01$). In this case, the associations between the 6 air pollutants related in this work were studied using a Spearman correlation test. Air pollutant distributions and their Spearman's *rho* coefficient ($\boldsymbol{r_s}$), are given in figure. 3. It can be observed that the correlations for each pair of the analyzed air pollutants were significantly positive (all p<0.01) except O_3 that had an inverse, but also statistically significant ($p < 0.01$) correlation with all the other pollutants. A weaker correlation between SO_2 and CO, NO_2, NO_x and O_3 was observed, as shown by the Spearman correlation

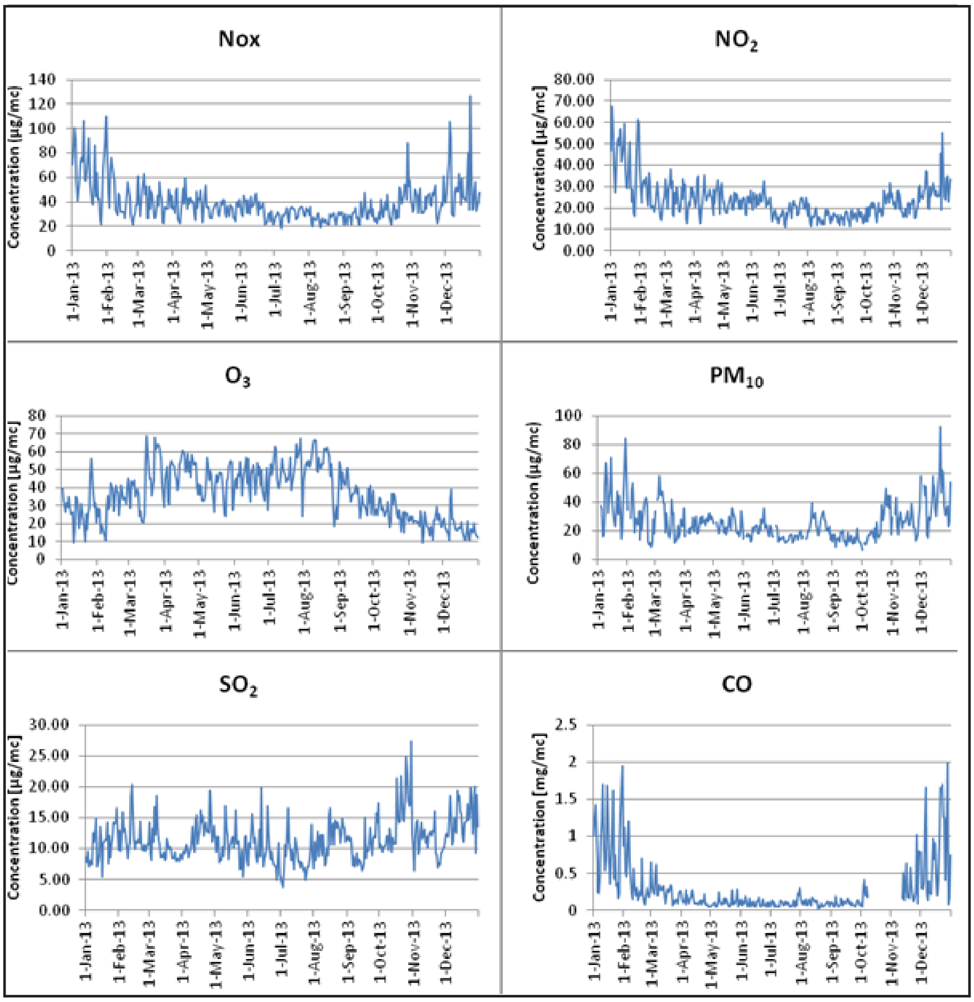


Fig. 1 Time series of air pollutants recorded at automated monitoring stations in Ploiesti during 2013 year (daily average values)

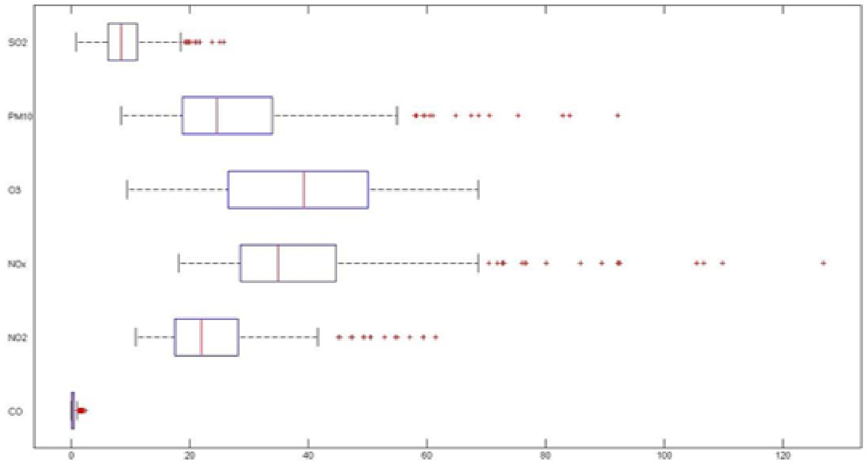


Fig.2. Box plots of daily average of concentrations for the analyzed pollutants (SO_2, PM_{10} , O_3, NO_x, NO_2, CO) in Ploiesti city for the year 2013

Table 3
DESCRIPTIVE STATISTICS OF THE DAILY AVERAGE CONCENTRATIONS OF AIR POLLUTANTS RECORDED IN PLOIESTI CITY IN 2013

Air pollutant	*SO2*	*NO2*	*NOx*	*CO*	*O3*	*PM10*
Concentrations Units	µg/mc	µg/mc	µg/mc	mg/mc	µg/mc	µg/mc
Mean	11.36	23.73	39.53	0.29	36.82	26.07
Median	10.94	21.95	34.95	0.14	36.73	23.01
CV[%]	30.17	38.11	40.96	126.38	39.16	49.82
Kurtosis	1.78	4.42	5.31	5.74	-0.97	3.20
Skewness	0.97	1.81	1.99	2.42	0.07	1.51
Minimum	3.84	10.93	18.07	0.03	9.35	6.59
Maximum	27.37	67.47	126.88	1.99	68.68	92.25
N	365	365	365	330	365	343

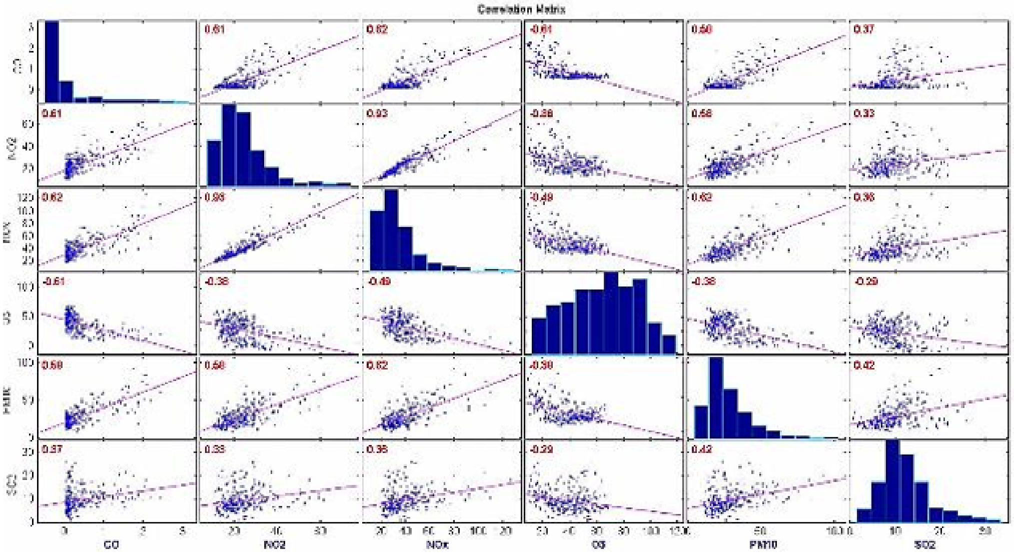


Fig. 3. Data distribution and Spearman rank correlation coefficient analysis of daily concentrations of air pollutants in Ploiesti city, recorded in 2013

coefficients (r_s = 0.372, r_s = 0.329, r_s = 0.357, *respectively* r_s = -0.285). A relatively good association occurred between PM_{10} and CO time series (r_s = 0.578).

Principal Component Analysis (PCA)

The Principal Component Analysis (PCA) was applied to determine the minimum set of factors that explains the variability of the analyzed air pollutants.

The analyzed data was stored in a n x m data matrix X where n represents the samples and m the variables. The data standardization was done by scaling the variable to have unit variance. We denote by Σ the covariance matrix of the standardized X data set. is a symmetric matrix with the eigendecomposition:

$$\Sigma = V\Lambda V^t,$$

V is the orthogonal eigenvector matrix, Δ is a diagonal matrix whose entries are the eigenvalues of Σ. The eigenvectors are ordered by eigenvalues which orders the components in order of significance.

In figure 4, it can be observed that the first 4 components explain 95.42% of the total variance. The first 2 eigenvectors defined as principal components (PC's) explain 78.35% of the total variance in data which means more than three quarters of the total variability in the standardized ratings, so that might be a reasonable way to reduce the dimensions.

The first component extracted in the PCA represents 63.88% amount of total variance of the observed variables and is correlated with some of the observed variables.

The second extracted component accounts for a maximal amount of variance in the data set that was not accounted for by the first component respectively 14.46% and is uncorrelated with the first component.

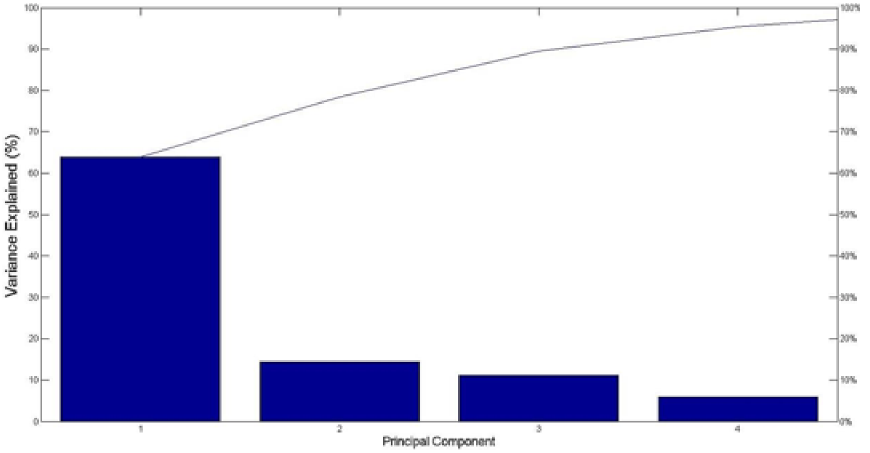


Fig. 4. Pareto chart of the principal components

Table 4

LOADINGS OF THE VARIABLES – DETERMINED PARAMETERS FOR THE FIRST PRINCIPAL COMPONENTS PC1 AND PC2

Loadings	CO	NO_2	NO_x	O_3	PM_{10}	SO_2
PC1	0.4630	0.4490	0.4715	0.3442	0.4380	0.2262
PC2	0.1280	0.2384	0.1568	0.0395	0.0276	0.9485

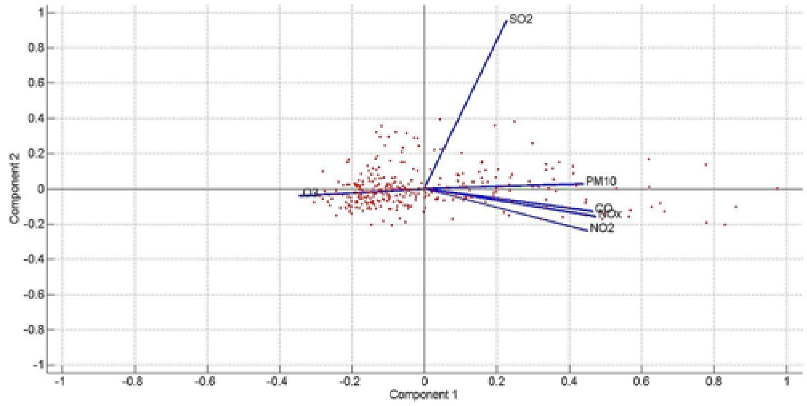


Fig. 5. Principal component scores for each observation and the loadings of each variable

Table 5

FORECASTED VALUES WITH ARIMA(3,1,3) AND OBSERVED VALUES AND THE CORRESPONDING ERRORS

Item	Forecast	Observed	Error	Relative error
CO	0.41	0.51	0.09	0.19
NO_x	27.61	27.15	0.46	0.01
NO_2	15.71	16.13	0.41	0.02
O_3	32.09	34.53	2.43	0.07
PM_{10}	19.41	22.89	3.47	0.15
SO_2	7.04	6.35	0.69	0.10

The principal component loadings values are presented in table 4 and the graphical representation in figure 5.

The first principal component is correlated with CO, NO_x, NO_2 and PM_{10}, which means that these elements vary together. If one increases, then the remaining components also increase. They are all positively related as they all have positive signs. The second principal component is strongly positive correlated with SO_2 almost identifying with it.

Air quality forecasting

In the present study ARIMA model was used to forecast the values of the six pollutants of concern. It were made more trials and the best results were obtain with ARIMA(3,1,3).

Due to the nature of our time series, it is recommended to logarithmate the data for the last 5 components. For exemplification, we tried to forecast the value in the day 261 using the first 260 observations. We used the observed values to evaluate the accuracy of the forecast.

For the data with significant variations recorded in Ploiesti such as CO, PM_{10}, and SO_2, the ARIMA statistical forecasting method did not provide satisfactory results the relative errors are bigger than 10%. Therefore, a more appropriate method may be the using of neural networks or the combination of the neural networks with wavelet decomposition of the data [14]. For the data corresponding to NO_x, NO_2, or O_3 time series, the relative error was 1%, 2% and 7%, respectively (as can be seen in table 5). Therefore, the ARIMA model provided satisfactory results.

Conclusions

Air pollution monitoring represents a problem with a constant increasing importance, nowadays. Statistics represents an important method in analyzing the air pollution and establishing the correlations between various elements. It also provides efficient tools for the forecasting pollution mechanisms based on time series. For some of the pollutants (NO_2, NO_x, O_3), the results were accurate enough. However, these forecasting statistical methods have their own limitations and do not produce satisfactory results for some pollutants (PM_{10}, CO) even if the forecasted period is short (one day ahead). More tests for establishing the suitability of ARIMA model in forecasting air pollutants for various time scales are required in conjunction with other conventional or artificial intelligence models.

Acknowledgments: This study received funding from the European Economic Area Financial Mechanism 2009 - 2014 under the project ROKIDAIR "Towards a better protection of children against air pollution threats in the urban areas of Romania" contract no. 20SEE/30.06.2014.

References

1. TRIVELLAS, T., HRISSANTHOU, V., Study of time series of air pollution in the city of Kavala, Greece, Proceedings of the 8th International Conference on Environmental Science and Technology, Lemnos Island, Greece, 8-10 September 2003, Vol. B, 2003, p. 806-814.

2. CAI, X.H., Time Series Analysis of Air Pollution CO in California South Coast Area, with Seasonal ARIMA model and VAR model, Los Angeles: University of California, 2008.

3. DUNEA, D., IORDACHE, S., Time series analysis of the heavy metals loaded wastewaters resulted from chromium electroplating process, Environmental Engineering and Management Journal*,* **10**(3), 2011, p. 421-434.

4. DUNEA, D., IORDACHE, S., Time series analysis of air pollutants recorded from Romanian EMEP stations at mountain sites**,** Environmental Engineering and Management Journal*,* **14**(11), 2015, p. 2725-2735

5.DUNEA, D., IORDACHE, S., ALEXANDRESCU, D., DINCA, N., Screening the weekdays/weekend patterns of air pollutant concentrations recorded in Southeastern Romania**,** Environmental Engineering and Management Journal, **13**(12), 2014, p. 3105-3114.

6. DUNEA, D., IORDACHE, S., RADULESCU, C., POHOATA, A., DULAMA, I., A multidimensional approach to the influence of wind on the variations of particulate matter and associated heavy metals in Ploiesti city, Romanian Journal of Physics, **61**(7-8), 2016, p. 1354-1368.

7.TOULOUMI, G., ATKINSON, R., TERTE, A.L., Analysis of health outcome time series data in epidemiological studies, Environmetrics, 15, 2004, p. 101-117.

8.DUNEA, D., IORDACHE, S., LIU, H-Y., BØHLER, T., POHOATA, A., RADULESCU, C., Quantifying the impact of PM2.5 and associated heavy metals on respiratory health of children near metallurgical facilities, Environmental Science and Pollution Research*,* **23**(15), 2016, p. 15395–15406.

9.*** Air Quality Plan for Prahova County 2015-2020 (*in Romanian*). Available at http://www.cjph.ro/Plan_Aer_2015-2020.pdf

10. DIAZ-ROBLES, L., ORTEGA J.C., FU J., REED G., CHOW, J.C., WATSON, J., MONCADA-HERRERA, J., A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile, Atmospheric Environment **42**, 2008, p. 8331–8340.

11. BOX G., JENKINS G., Time series analysis: Forecasting and control, San Francisco: Holden-Day, 1970.

12.SIEW, L.Y., CHIN, L.Y., WEE,P., ARIMA and integrated ARFIMA models for forecasting air pollution index in Shah Alam, Selangor, The Malaysian Journal of Analytical Sciences, **12**(1), 2008, p. 257-263 .

13. ZAFRA, C., ANGEL, Y., TORRES, E., ARIMA analysis of the effect of land surface coverage on PM_{10} concentrations in a high-altitude megacity, Atmospheric Pollution Research, 2017, In Press, http://dx.doi.org/10.1016/j.apr.2017.01.002.

14. DUNEA, D., POHOATA, A., IORDACHE, S., Using wavelet-feedforward neural networks to improve air pollution forecasting in urban environments., Environ. Monit. Assess., 2015, 187(7):477

Manuscript received: 110.01.2017