Mathematical Models for Inorganic Pollutants in Constanta Area, Romania

ALINA BARBULESCU, LUCICA BARBES*
Ovidius University of Constana, 124 Mamaia Blv., 900527, Constana, Romania

To develop analytical models for main inorganic pollutants (SO_2, NO_2) from risk areas in Constana district, data have been collected in the period January 2008 – June 2009, using Nitrogen Oxide Analyzer and Sulphurous Hydrogen Analyzer. Models for the evolution of pollutant's concentrations have been determined by GRNN and Box-Jenkins methods. For ten of the twelve studied series, the accuracy of GRNN models was very high (over 90%). For the rest, AR models that improved the fitting quality have been determined.

Keywords: SO_2, NO_2, GRNN, AR

The presence and persistence in atmosphere of some inorganic pollutants like NO_x, H_2S-SO_2, CO and other particulate matter (PM_{10} or $PM_{2.5}$) pose severe problems for the human health. The basic air quality management and the control of health risks include the continuous monitoring of pollutants concentrations as well as studies performed to characterize the atmospheric emissions or imissions, the physical and chemical processes and the pollutants dissipation. The pollution directly affects the human health and produces environmental modification [8, 11]. The pollution sources are mainly divided into two categories: (a) natural pollution, such as NO and NO_2 from bacteria activities, (b) artificial air pollution, due to the human activities, such as industrial activities, transportation or domestic/household combustion. The second one has as results pollutants as CO from methane (CH_4) incomplete oxidation, unburned hydrocarbons, H_2S and CH_4 from anaerobic decomposition of organic compounds etc.

Quality air regulations have been legislated in many countries, but there is no concordance between the values given by them. Therefore, it is necessary the establishment of maximum admissible concentrations of some pollutants, accepted at the world level. For example, the maximum annual admissible concentration of SO_2 in atmosphere is of 0.25 mg/ m³ in Romania, Suedia, Israel; 0.275 mg/m³ in Danemarca and 0.3 mg/m³ in Canada and Turkey, differing from the values accepted in Poland - 0.35 mg/m³, USA - 0.365 mg/m³ and Switzerland and Russia - 0.5 mg/m³.

The analysis of pollutants concentration has been done in many regions of Romania, in 2010. The report presented in [14] emphasizes that there were urban areas where the pollutants maximum admissible concentrations were exceeded.

Considering the importance of preserving the air quality for the human health, the aim of our work was the monitoring of the concentrations of emissions of H_2S-SO_2, NOx, CO and PM_{10} from Constanta district (three sites – Gate 1, Gate 3 and Epuration station in Constanta city, two in Navodari area: Navodari_Ind (industrial area) and Navodari_Camp and one in Corbu village) for 18 months from January 2008 to June 2009 and modelling the data. The monitoring sites were setup in compliance with the European air quality legislation.

The charts of mean monthly concentrations of SO_2 and NO_2 are given in figure 1. The maximum values of emission rates were respectively: 216.66 μg/m³ - for hydrogen sulfide and sulphur dioxide, 80.8 μg/m³ - for nitrogen oxides, 6.612 mg/m³ - for carbon monoxide and 42 μg/m³ for particle matter with an aerodynamic diameter of 10 μm or less.

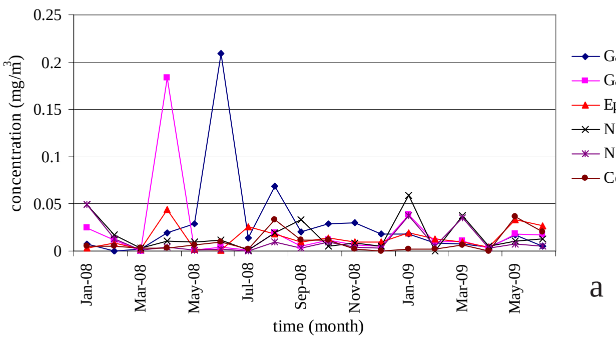
It was observed that in summer the emission rates were lower than in winter because the sea has a strong influence on climate through the storage and transport of particle matter or pollutants.

In this article we present only the models of evolution of SO_2 and NO_2 concentrations.

Methodology

To collect the data concerning the concentration of SO_2 and NO_2, specific measurements have been done, using a mobile laboratory.

SO_2 concentration has been determined using Thermo Scientific Hydrogen Sulfide and Sulfur Dioxide Analyzer,



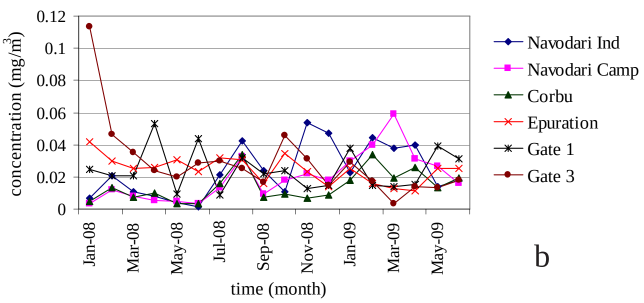


Fig. 1. Evolution of mean monthly concentration of: (a) SO_2; (b) NO_2

* *email: lbarbes@univ-ovidius.ro; Tel.: 0241606488*

Model 450i (Thermo Environmental Instruments) that utilizes pulsed fluorescence technology. The instrument operates on the principle that H_2S can be converted to SO_2. As the SO_2 molecules absorb ultraviolet (UV) light and become excited at one wavelength, the molecules then decay to a lower energy state emitting UV light at a different wavelength. The pulsing of the UV source lamp serves to increase the optical intensity whereby a greater UV energy throughput and lower detectable SO_2 concentration are realized [16].

NO_2 concentration has been measured by Thermo Environmental 42C Nitrogen Oxides (NO-NO_2-NO_x) analyzer. It is based on the principle that nitric oxide and ozone react to produce a characteristic luminescence with intensity linearly proportional to the NO concentration. Infrared light emission results when electronically excited NO_2 molecules decay to lower energy states. Nitrogen dioxide (NO_2) must first be transformed into NO, before it can be measured using the chemiluminescent reaction. The NO and NO_x concentrations calculated in the NO and NO_x modes are stored in memory. The difference between the concentrations is used to calculate the NO_2 concentration [16].

To determine some appropriate models for the evolution of pollutants' concentration in the study area, a combination of statistical techniques, Box – Jenkins and heuristic methods has been used. The stages of our study were:

1. Performing statistical analyses in order to determine the outliers' presence (using the boxplot), the series autocorrelation (using the autocorrelation function) and the presence of a linear trend (using the Mann Kendall test and Sen' slope estimate) [15];

2. Determining a model, using Generalized Regression Neural Networks (GRNN) or a Box Jenkins procedure, for the initial series, if no trend was determined, or for the series obtained after the trend removal, otherwise.

In the following we shall shortly remember some basic notions related to the procedures used for the statistical analysis and data modeling.

A discrete process (a time series) in time is a sequence of random variables $(X_t)_{t\in N^*}$ (shortly denoted by (X_t)). The autocorrelation of a random process describes the correlation between values of the process at two different points in time, as a function of the two times or of the time difference. If (X_t) is a time series, the autocorrelation function of (X_t) at lag h (h ∈ N)* is defined by: $\rho(h)=\dfrac{\gamma(h)}{\gamma(0)}$, $h\in \mathrm{N}^*$, where $\gamma(h)=Cov(X_t, X_{t+h})$ is the covariance of X_t and (X_{t+h}).

If x_1, ... , x_n are observed data, from the previous formula we obtain the empirical autocorrelation function, denoted by ACF. Together with ACF the confidence limits at the confidence level of 95% or 99% are also determined. If for the selected sample, the values of ACF are inside the corresponding confidence interval, we conclude the series is not auto-correlated.

(X_t) is said to be stationary if it has a finite mean and its covariance depends only on the lag between two points in the series. A stationary process (X_t) is called a white noise if X_t, X_{t+h} are uncorrelated for every $h \neq 0$ identically distributed, with null expectance, and the same variance.

A discrete process is said to be *autoregressive of p order* – AR(p) – if:

$$X_t=\sum_{i=1}^{p}\varphi_i X_{t-i}+Z_t, \forall\ t\in \mathrm{N}^*,\ \varphi_p\neq 0,$$

where (Z_t) is a white noise [3].

The Mann-Kendall test is used to test the existence of a monotonic trend of a time series with no seasonal or other cycle. The data need not to conform to any particular distribution. This test is applicable in cases when the data values x_i of a time series can be assumed to obey the model:

$$x_i = f(t)+\varepsilon_i, \qquad (1)$$

where *f(t)* is a continuous monotonic increasing or decreasing function of time and the residuals ε_i can be assumed to be from the same distribution with zero mean. The null hypothesis in this test is: H_0 "the observations x_i are randomly ordered in time", and its alternative is: H_1: "there is a monotonic trend".

In the hypothesis that $x_i \neq x_j$, $\forall i\neq j$, the test statistic is: $Z=\dfrac{\dfrac{4p}{n(n-1)}-1}{\sqrt{\dfrac{2n(n+5)}{9n(n-1)}}}$, where p is the number of occurrences of $x_i < x_{ij}$, for i < j . For a given level of significance, α, the null hypothesis is rejected if $|Z| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is obtained from the standard normal cumulative distribution tables. More specifically, $Z > z_{1-\alpha/2}$ indicates the existence of an increasing trend and $Z < z_{1-\alpha/2}$ implies a decreasing one [6].

The Sen's method is used if the function *f* in eq.(1) is linear:

$$f(t) = Qt + B$$

The slope, *Q*, is determined to be the median of all the data values pairs, x_i, x_j, with $i > j$.

This method is robust to the outliers presence [15].

An *outlier* is a value that appears to deviate markedly from other members of the sample in which it occurs [1]. The standard scores may be used to detect outliers in the hypothesis that a variable is normally distributed. When a variable is not normally distributed, a boxplot may be more effective in identifying outliers. Cases with values between 1.5 and 3 box lengths (inter-quartile range) from the upper or lower edge of the box are identified as outliers.

A neural network (NN) is a mathematical model for solving different problems. There are two types of NN: feed-forward and recurrent. In the feed-forward NN, used in the present work, the information is propagated from the input to the output. GRNN belongs to the nonparametric kernel regression models. Its structure consists of four layers: input layer, pattern layer, summation layer and output layer as shown in figure 2.

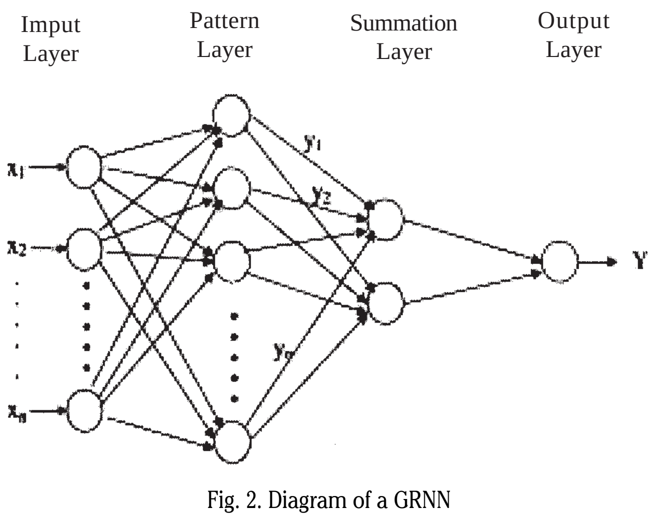


Fig. 2. Diagram of a GRNN

Station	Minimum	Maximum	Mean	Std. Dev	Skewness	Kurtosis
Gate1	0	0.2097	0.0282	0.04793	3.573	13.702
Gate3	0	0.183	0.02055	0.041725	3.868	15.674
Epuration	0.0008	0.04435	0.014183	0.011858	1.193	1.062
Navodari_Camp	0.0001	0.0492	0.010769	0.014402	1.891	2.493
Navodari_Ind	0.0001	0.0585	0.016761	0.016921	1.447	1.205
Corbu	0	0.0365	0.009144	0.010687	1.794	2.514

Table 1
BASIC STATISTICS OF SO_2 CONCENTRATION (mg/m^3)

The input layer contains one neuron for each predictor variable. The input neurons standardize the range of the values by subtracting the median and dividing by the interquartile range. The pattern layer contains one neuron for each case in the training data set, and stores the predictor variables values and target values. When a new vector is entered into the network, it is subtracted from the stored vector representing each cluster center. Either the squares or the absolute values of the differences are summed and fed into a nonlinear activation function [12]. The activation function normally used is the exponential, but there are also other possible choices for it. Each pattern layer unit is linked to the two neurons in the summation layer: the denominator summation neuron, which adds up the weight values coming from each of the neurons in the pattern layer, and the numerator summation neuron, that adds up the weight values multiplied by the actual target value for each hidden neuron. The output layer, formed by one neuron, divides the values accumulated in the summation layer.

The neural network approach has been chosen for modeling our data since it has been proved that they gave good results in different prediction problems [4], [7]. The last years the neural networks has been extensively used since their performance is generally superior to those of the traditional statistical methods, as multiple regression, autoregressive models and classification trees [5, 13].

Results and discussions

Results concerning the SO_2 concentration evolution

Table 1 contains the basic statistics of SO_2 concentration series. The series distributions are not homogenous, as we remark from the values of skewness and kurtosis, whose values varies from 1.193 to 3.868, respectively from 1.205 to 15.674. Remember that the skewness is a measure of the asymmetry of the probability distribution, a positive skew indicating that the tail on the right side is longer than on the left side. In this case, the heaviest tail corresponds to the distribution of Gate 1. Kurtosis is a descriptor of the shape of a probability distribution: in this case, all the distributions are leptokurtic, the highest peak corresponding to Gate 3.

Also, the series are not correlated, nor normally distributed and they contain outliers that can affect the models goodness of fit. In figure 3 the outliers are represented by numbered small circles; the numbers give the position of the data in the series, knowing that 1 denotes January 2008 and 18, June 2009.

The results of the Mann - Kendall test, with Sen's slope estimate, are presented in figure 4, where: the dots represent the concentration values, the black line is Sen'slope estimate, the grey curve is the residual and the dashed lines are the limits of the confidence intervals at the confidence levels of 99% and 95%.

Analyzing the chart and comparing the values of Z statistic with $z_{0.975} = 1.96$, we conclude that there is no linear monotonic trend for the series. Therefore, the GRNN models have been built for the initial series. The models are very good for all the series, but Navodari_Camp, because the proportion of variance explained by model and the correlation between the actual and predicted values are very close to 1, the normalized mean square error is very close to 0 and the variation coefficient is also very small (table 2).

The charts from figure 5 (a)-(e) present the actual data, the predicted ones and the forecast for three months for the series, but Navodari_Camp. The actual and predicted curves are practically superposed in figure 5 (a), (d), (e). In figure 5(f) we remark the errors in the model for Navodari_Camp, that confirms the results from table 2, column 6.

To improve the modelling result for Navodari_Camp series, the mean has been subtracted from the data, and an AR(1) model has been built:

$$X_t = -0.527X_{t-1} + \varepsilon_t, \qquad (2)$$

(ε_t) being a white noise with the standard deviation of 0.00023 and the correlation between the calculated and actual values of 85% (fig. 6).

Results concerning the NO_2 concentration evolution

The distributions of NO_2 concentrations are less asymmetric than those of SO_2, three are platycurtic and three leptokurtic (table 3). They do not present outliers,

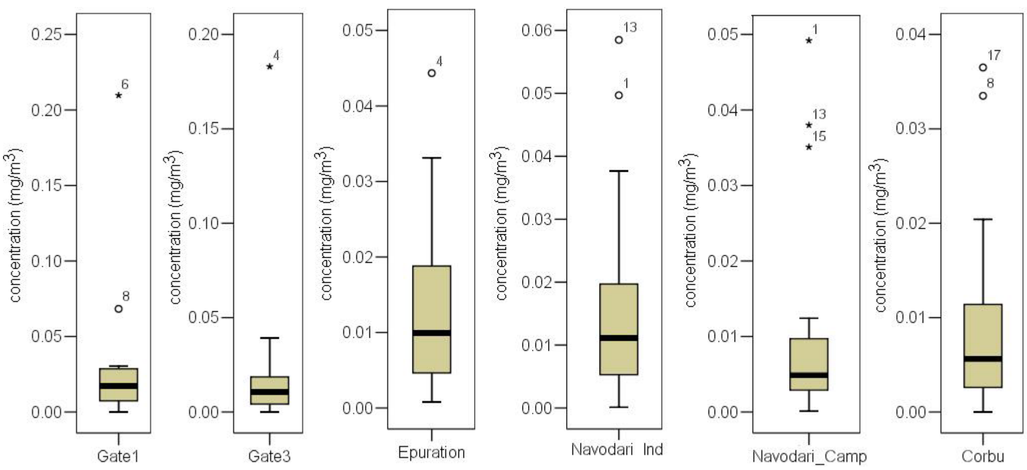


Fig. 3. Boxplots of SO_2 series

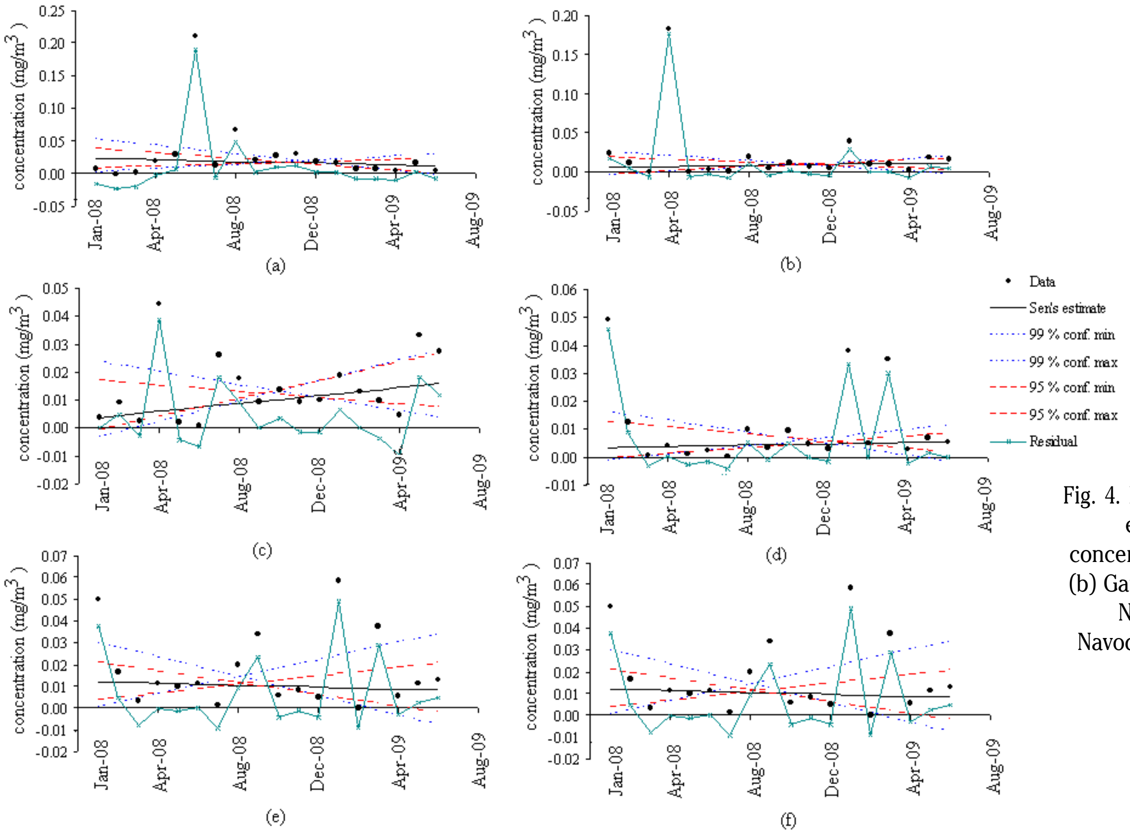


Fig. 4. Results of Sen's slope estimate for SO_2 concentration at: (a) Gate 1, (b) Gate 3,(c) Epuration, (d) Navodari_Ind, (e) Navodari_Camp, (f) Corbu

Table 2
QUALITY INDICATORS FOR GRNN MODELS FOR SO_2 CONCENTRATION

	Gate 1	Gate 3	Epuration	Navodari_Ind	Navodari_Camp	Corbu
Proportion of variance explained by model (%)	99.776	97.664	99.136	100	61.272	99.931
Coefficient of variation	0.078	0.030	0.075	0.0002	0.808	0.029
Normalized mean square error	0.002	0.023	0.008	0.0000	0.387	0.0007
Correlation between actual and predicted	0.9989	0.988	0.995	1.0000	0.798	0.999

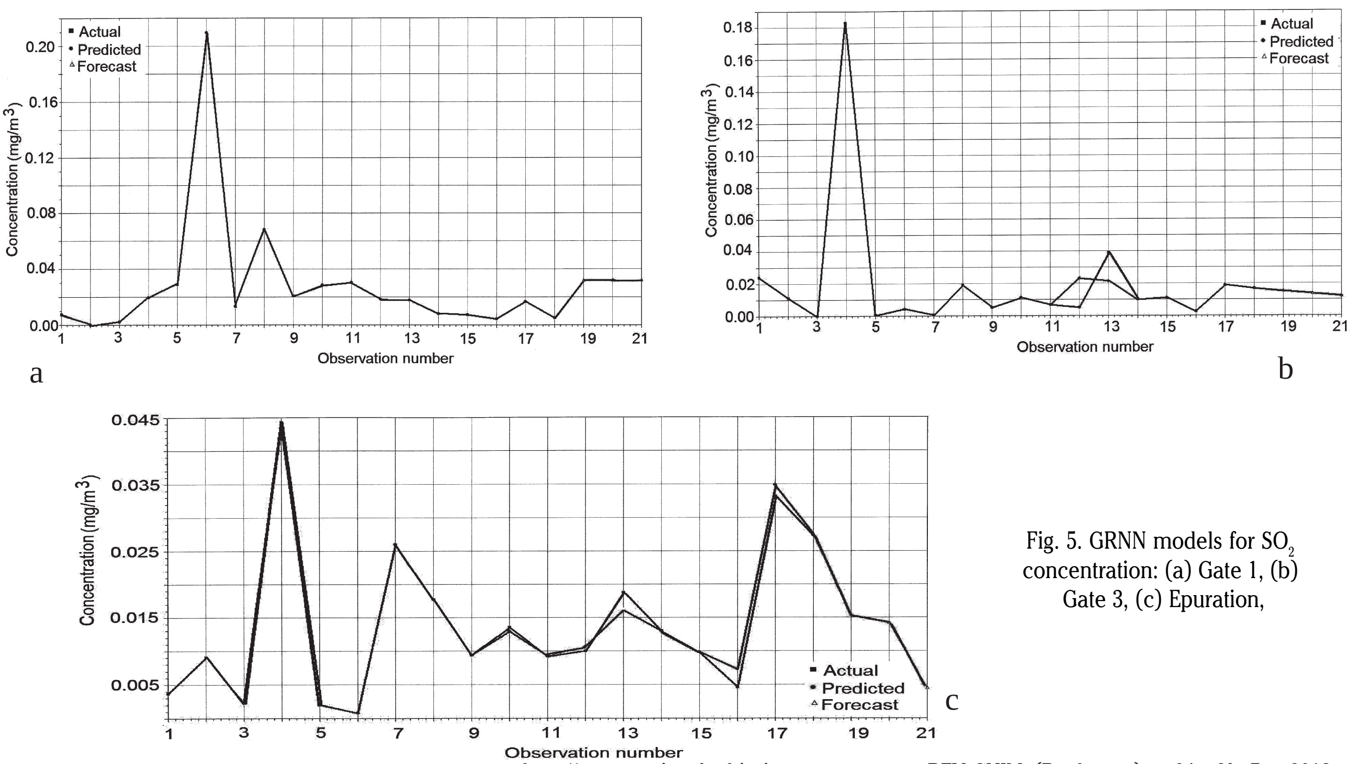
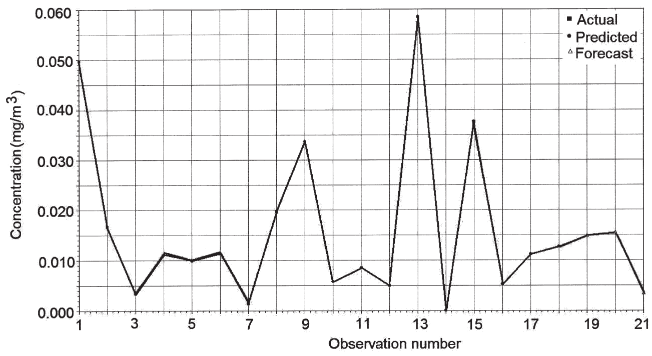
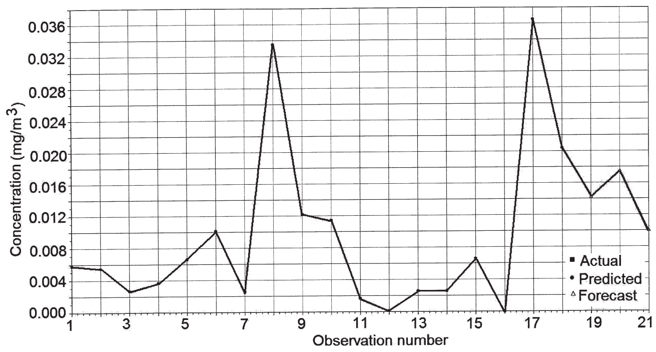


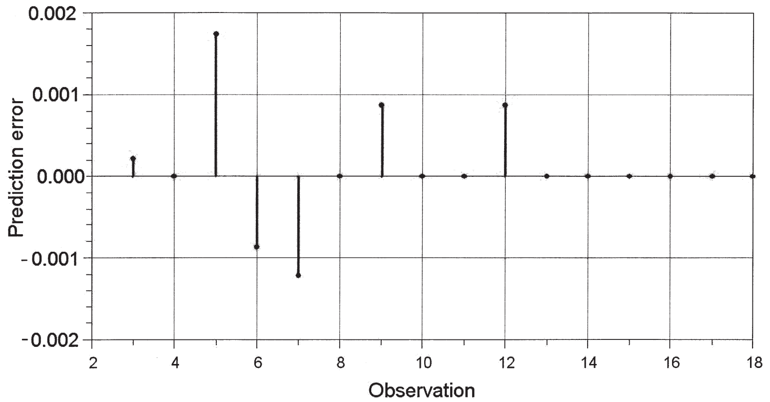
Fig. 5. GRNN models for SO_2 concentration: (a) Gate 1, (b) Gate 3, (c) Epuration,



d



e



f

Fig. 5. GRNN models for SO_2 concentration: (d) Navodari_Ind, (e) Corbu, (f) Model errors for Navodari Camp series

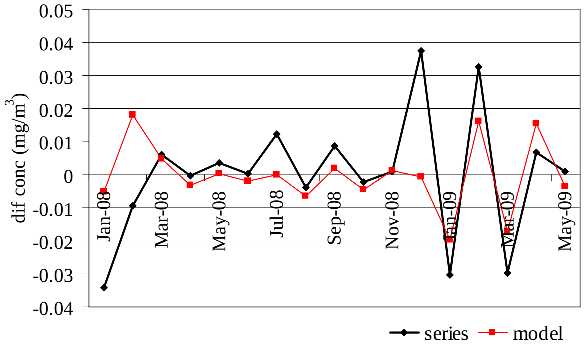


Fig.6. AR model for difference of first order of Navodari_Camp series

Station	Minimum	Maximum	Mean	Std. Dev	Skewness	Kurtosis
Gate1	0.0083	0.0528	0.0242	0.0126	0.795	-0.134
Gate3	0.0036	0.1137	0.0291	0.0238	2.829	9.825
Epuration	0.0110	0.416	0.0245	0.0081	0.035	-0.290
Navodari Camp	0.0032	0.0587	0.0195	0.0147	1.136	1.404
Navodari_Ind	0.0014	0.0534	0.0236	0.0164	0.432	-1.189
Corbu	0.0032	0.0341	0.0140	0.0095	1.019	0.307

Table 3
BASIC STATISTIC OF NO_2 CONCENTRATION (mg / m³)

	Kolmogorov-Smirnov[(a)]			Shapiro -Wilk		
	Statistic	df	p -val	Statistic	df	p -val
Gate1	.155	18	0.200[(*)]	0.927	18	0.174
Gate3	.244	18	0.006	0.697	18	0.000
Epuration	.159	18	0.200[(*)]	0.951	18	0.443
Navodari_Camp	.150	18	0.200[(*)]	0.905	18	0.071
Navodari_Ind	.169	18	0.185	0.920	18	0.130
Corbu	.160	18	0.200[(*)]	0.888	18	0.035

* This is a lower bound of the true significance.
[a] Lilliefors Significance Correction

Table 4
NORMALITY TESTS

Series	Z	Q	Qmin99	Qmax99	Qmin95	Qmax95
Gate 1	-0.08	0.000	-0.002	0.001	-0.001	0.001
Gate 3	-3.03	-0.002	-0.004	0.000	-0.003	-0.001
Epuration	-2.20	-0.001	-0.002	0.000	-0.002	0.000
Navodari Camp	2.95	0.002	0.000	0.003	0.001	0.003
Navodari Ind	1.74	0.001	-0.001	0.004	0.000	0.003
Corbu	2.24	0.001	0.000	0.002	0.000	0.002

Table 5
RESULTS OF MANN KENDALL AND SEN'S SLOPE ESTIMATE FOR NO_2 CONCENTRATION SERIES

excepting Gate 3 (for which the outlier is the first value of the series). The concentrations of NO_2 are higher than those of SO_2.

At the significance level of 0.01 we can not find enough evidence to reject the hypothesis that the data are normally distributed, for all series, but Gate 3, as it result from table

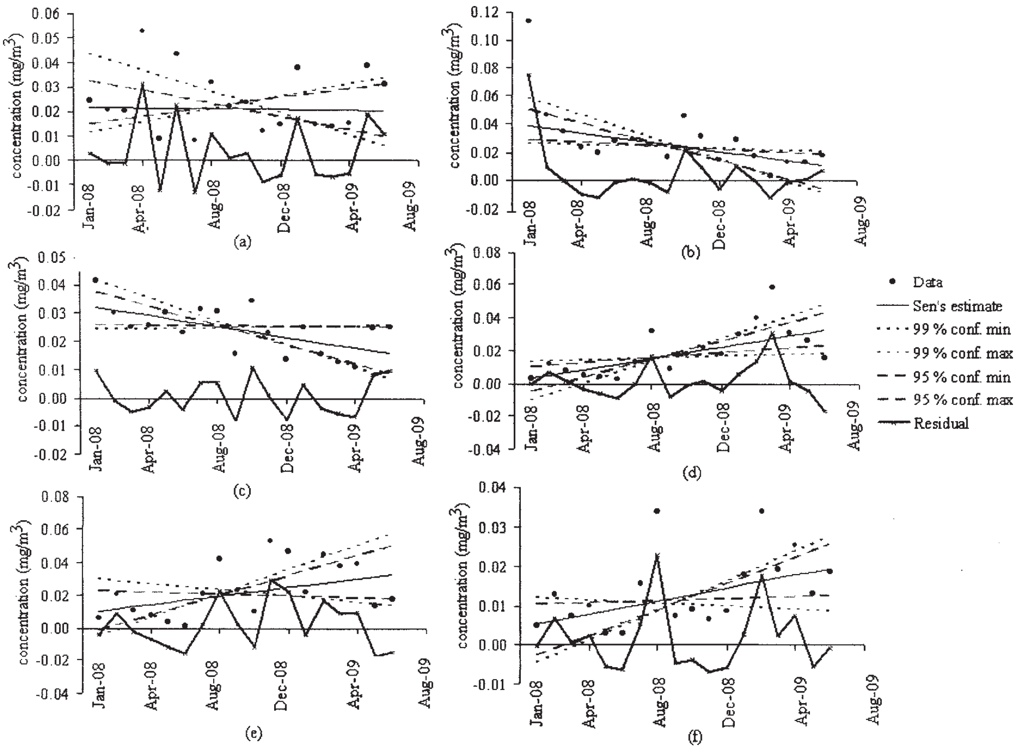


Fig. 7. Results of Sen's sope estimate for NO_2 concentration at: (a) Gate 1, (b) Gate 3, (c) Epuration, (d) Navodari_Ind, (e) Navodari Camp, (f) Corbu

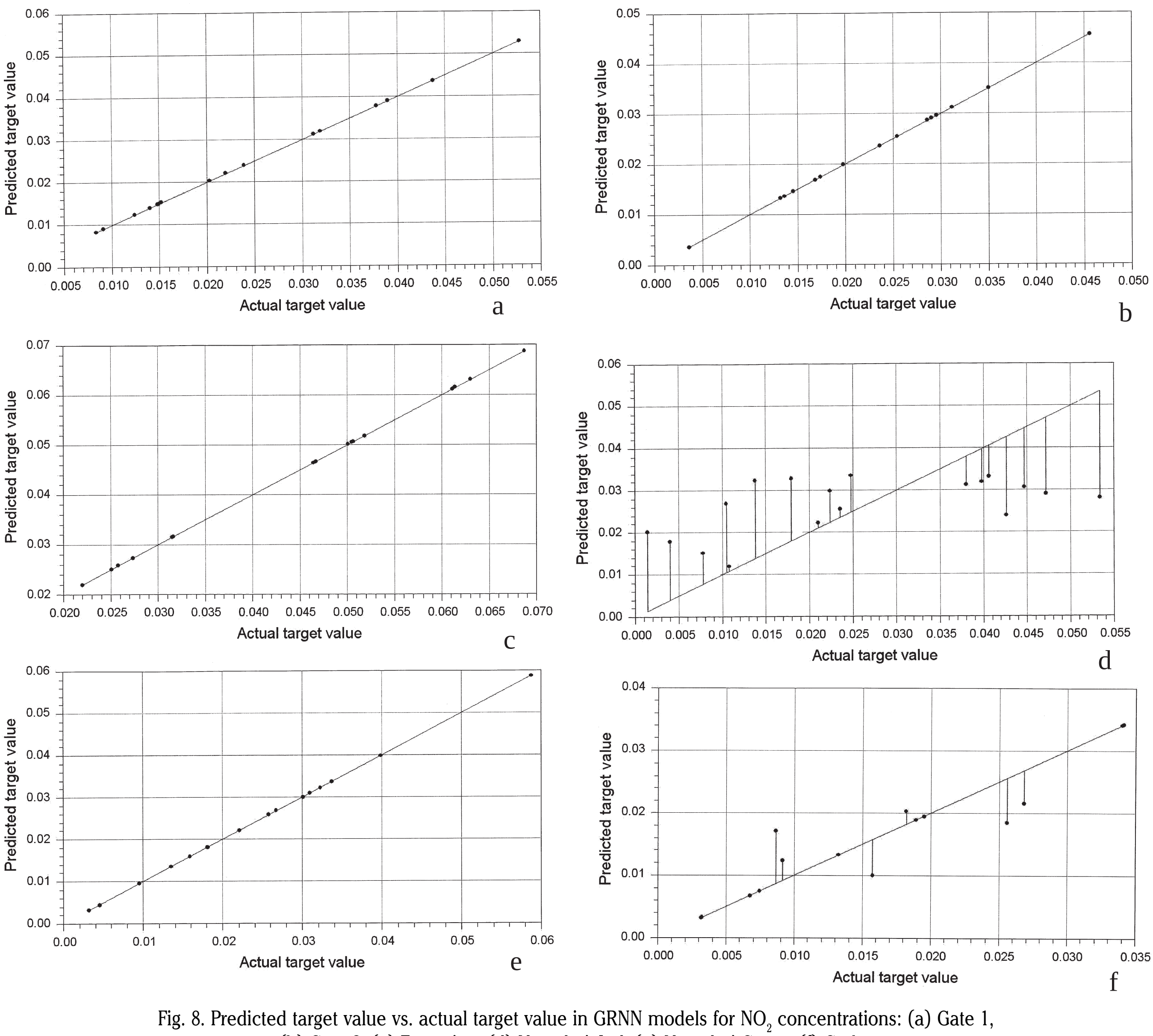


Fig. 8. Predicted target value vs. actual target value in GRNN models for NO_2 concentrations: (a) Gate 1, (b) Gate 3, (c) Epuration, (d) Navodari_Ind, (e) Navodari Camp, (f) Corbu

	Gate 1	Gate 3	Epuration	Navodari_ Ind	Navodari_ Camp	Corbu
Proportion of variance explained by model (%)	99.868	99.871	100	30.285	98.261	84.201
Coefficient of variation	0.018	0.028	0.000	0.531	0.089	0.254
Normalized mean square error	0.001	0.001	0.000	0.697	0.017	0.158
Correlation between actual and predicted	0.999	0.999	100	0.550	0.993	0.919

Table 6
RESULTS OF GRNN MODELS

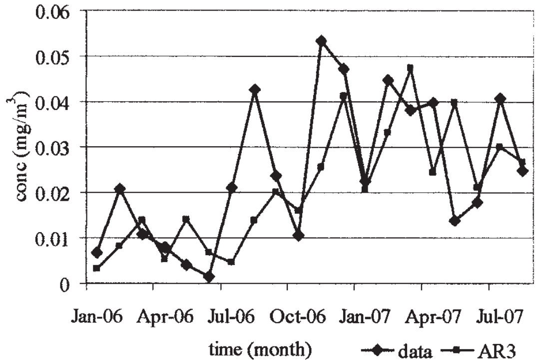


Fig. 9. MA(3) model for Navodari_Ind series

4, where the results of Kolmogorov – Smirnov and Shapiro – Wilk are presented and where df represents the degree of freedom and p-val is the p – value. We remember that if the p- value is bigger than the fixed significance level (0.01, in this case), the null hypothesis (the normality, in our case) can not be rejected.For a significance level $\alpha=0.05$, the results of Mann – Kendall test and Sen's slopes estimates are the presented in table 5 and figure 7. In table 5, Z is the value of Mann – Kendall test statistic, Q is the slope estimate, Qmin99, Qmin95, respectively Qmax99 and Qmax95 are the lower and the upper limits of the confidence interval of Q, at the confidence levels of 99% and 95%. Since 2.95 and 2.24 > $z_{0.975} = 1.96$, we deduce that there is a linear increasing trend for Navodari_Camp and Corbu series. Since -3.03 and -2.20 < $z_{0.025} = -1.96$, it results that there is a linear decreasing trend for Gate 3 and Epuration series. Since -1.96 < -0.08, 1.72 < 1.96, there is no linear trend for Navodari_Ind and Gate 1 series.

Analogous to the case of the SO_2series, the GRNN models' quality for NO_2 concentration series can be remarked after the study of figure 8 and table 6. The only case when the model did not fit well the data is Năvodari_Ind series, since the proportion of variance explained by model was only of 30.285%. The previous assertions are sustained by figure 8, where the predicted value vs the actual ones are presented. We also remark the existence of some big deviations of the predicted errors from the actual ones in the case of Corbu series, confirming the correlation between the estimated and actual values of only 0.919.

To improve the result obtained for Navodari_Ind series, after making the first order differences of data and the mean subtraction, an AR(3) model has been determined:

$$X_t = 0.5571X_t - 0.2206X_{t-2} + 0.581X_{t-3} + \varepsilon_t, \quad `(3)$$

where (ε_t) is a white noise with the variance 0.000178 (fig. 9), the correlation between the predicted and the actual values being 0.71.

Conclusions

In this article we presented models for the NO_2 and SO_2 pollutants' series evolution in the district of Constanta, at six sites. The GRNN approach was appropriated since for 10 of 12 series we obtained good models, with the correlation between actual values and predicted ones over 90%. For the other series, AR models have been built, improving the GRNN models. Even if classical methods are not all the time suitable to model phenomena with high variability, Box-Jenkins methods gave good results in the study cases.

References

1. BARNETT, V., LEWIS, T., Outliers in Statistical Data, 3rd edition, John Wiley & Sons, 1994
2. BĂRBULESCU, A., BARBES, L., Models for Pollutants Evolution in an Urban Area, Mathematical Models and Methods in Applied Sciences, 2012, p. 52.
3. BĂRBULESCU, A., BĂUTU, E., Alternative models in precipitation analysis, An. 't. Univ. Ovidius, Math, vol.**17**(3), 2009, p. 45.
4. GRIVAS, G., CHALOULAKOU, A., Artificial neural network models for prediction of PM_{10} hourly concentrations, in the Greater Area of Athens, Greece, Atmospheric Environment, vol. **40**, no. 7, 2006, p. 1216.
5. IBARRA-BERASTEGI, G., ELIAS, A., BARONA, A., SAENZ, J., EZCURRA, A., De ARGANDO A, J. D., From diagnosis to prognosis for forecasting air pollution using neural networks: air pollution monitoring in Bilbao (Spain), Environmental Modeling & Software, **23** (5), 2008, p. 622.
6. KENDALL, M. G., Rank Correlation Methods, Griffin, London, 1955.
7. PÉREZ-ROA, R., CASTRO, J., JORQUERA, H., PÉREZ-CORREA J.R., VESOVIC, V., Air - pollution modelling in an urban area: Correlating turbulent diffusion coefficients by means of an artificial neural network approach, Atmospheric Environment, **40**, 2006, p. 109.
8.SALCEDO, D., CASTRO, T., RUIZ-SUÁREZ, L.G., GARCÍA-REYNOSO, A., TORRES-JARDÓN, R., TORRES-JARAMILLO, A., MAR-MORALES, B.E., SALCIDO, A., CELADA, A.T., CARREÓN-SIERRA, S., MARTÍNEZ, A.P., FENTANES-ARRIAGA, O.A., DEUSTÚA, E., RAMOS-VILLEGAS, R., RETAMA HERNÁNDEZ, A., SAAVEDRA, M.I., SUÁREZ-LASTRA, M., Study of the regional air quality south of Mexico City (Morelos state), Science of the Total Environment, **414**, 2012, p. 417
9. SALMI, T., MÄÄTTÄ, A., ANTTILA, P., RUOHO-AIROLA, T., AMNELL, T., Detecting trends of annual values of atmospheric pollutants by the Mann-Kendall test and Sen's slope estimates –the Excel template application MAKESENS, Finnish Meteorological Institute, Air Quality Research, Helsinki, 2002.
10. SHAPIRO, S. S., WILK, M. B., An Analysis of Variance Test for Normality (Complete Samples), Biometrika, Vol. **52**, No. 3/4, 1965, p. 591.
11. SINGH, K.P., GUPTA, S., KUMAR, A., PRASAD SHUKLA, S., Linear and nonlinear modeling approaches for urban air quality prediction, Science of The Total Environment, **426**, 2012, p. 244
12.SPRECHT, D. F., A General Regression Neural Network, IEEE Transactions on Neural Networks, vol. **2**, no. 6, 1991, pp. 568.
13.WANG, J.S., CHAN, T.L., CHEUNG, C.S., LEUNG, C.W., HUNG W.T., Three-dimensional Pollutant Concentration Dispersion from a Vehicular Exhaust Plume in the Real Atmosphere, Atmospheric Environment, **40**(3), 2006, p. 484.
14. *** www.anpm.ro/Files/cap13_200912174933391.pdf
15. *** www.emep.int/assessment/MAKESENS_1_0.xls
16.***www.thermo.com/eThermo/CMA/PDFs/Various/156File_17809.pdf

Manuscript received: 19.12.2012