MULTIPLE LINEAR REGRESSION MODELS BASED ON METEOROLOGICAL PARAMETERS FOR FORECASTING OF NITROGEN OXIDES CONCENTRATIONS LEVELS

Marko Uzunov

University of Chemical Technology and Metallurgy – Sofia
e-mail: marko.uzunov@gmail.com

***Keywords**: Nitrogen oxide, nitrogen dioxide, regression model, carbon monoxide, meteorological factors*

***Abstract:** The investigation deals with prediction of the air pollution of Sofia city with nitrogen oxides by applying a multiple linear regression approach.The models were designed based on the data for four meteorological factors: ambient temperature,air humidity, wind speed, sun radiation and content of the carbon monoxide used as additional factor. Data for the period April – June 2017 were used. The models possess index of agreement between 0.8498 and 0.9635; prediction accuracy from 0.7299 to 0.8681, coefficient of determination between 0.7287 and 0.8671 and normalized absolute error from 0.1893 to 0.3735. Most of the models enable predicting the nitrogen oxides pollution with an error below 10 %.*

МНОГОЛИНЕЙНИ РЕГРЕСИОННИ МОДЕЛИ, БАЗИРАНИ НА МЕТЕОРОЛОГИЧНИ ФАКТОРИ ЗА ПРОНОЗИРАНЕ КОНЦЕНТРАЦИОННИТЕ НИВА НА АЗОТНИТЕ ОКСИДИ

Марко Узунов

Химикотехнологичен и металургичен университет – София
e-mail: marko.uzunov@gmail.com

***Ключови думи:** Азотен монооксид, азотен диоксид, регресионен модел, въглероден монооксид, метеорологични фактори*

***Резюме:** Изследването е свързано с прогнозиране замърсяването на въздуха на град София с азотни оксиди чрез прилагане на множествена линейна регресия. Моделите бяха разработени въз основа на данните за четири метеорологични фактора: температура на околната среда, влажност на въздуха, скорост на вятъра, слънчева радиация и съдържание на въглероден оксид като допълнителен фактор. Използвани са данни за периода април – юни 2017 г. Моделите притежават индекс на съгласие между 0.8498 и 0.9635; точност на прогнозиране от 0.7299 до 0.8681, коефициент на детерминация между 0.7287 и 0.8671; нормализирана абсолютна грешка от 0.1893 до 0.3735. Повечето от моделите осигуряват прогнозиране на замърсяването с азотни оксиди с грешка под 10 %.*

Introduction

The problem concerning the air quality in the big cities has become so severe, that effective and timely information about changes in the level of air pollutions is urgently needed.

Nitrogen oxides belong to the most hazardous pollutants. The group of nitrogen oxides includes nitrogen in different valence states, but the harmful emissions mostly consist of NO and NO_2.

Nitrogen oxides are released into the atmosphere mainly in the form of N (II), which reacts with ozone thereby forming NO_2. The main source of NO and NO_2 are petroleum distillates used in automotive engines, and the road transport is the reason for approximately half of the exhausts in Europe's atmosphere. Therefore, the highest concentrations of nitrogen oxides are registered in the urban areas whit overloaded traffic. The thermal power stations and the industrial enterprises are also among the biggest pollutants with nitrogen oxides.It has been found that increased levels of nitrogen

oxides are usually produced in urban zones under stable weather conditions. There are a number of studies related to the impact of different meteorological factors on the concentration of nitrogen oxides, aiming at creating models for predicting the pollutants concentration [1–4]. Some authors state that the most important factors influencing the concentration of nitrogen oxides are the speed and direction of the wind [5].

The development of effective models for forecasting the air cleanness of the urban areas has drawn the attention of researchers nowadays. Although there are many proposed forecasting models and some of them are in use, it is still necessary to develop more accurate and simpler models [6, 7].

The prediction of the hourly concentrations of the pollutants according to data only for meteorological factors, available online in the weather forecast would be of real practical interest. Such approach would ensure preliminary information about the air pollution with the harmful nitrogen oxides.

The aim of this paper is to present the results of the application of MLR analysis in predicting NO_x concentration as the function of four meteorological parameters. In order to account for the influence of the traffic intensity on the NO_x content, one pollutant (CO), which is being emitted as a result of incomplete combustion of the fuels was also included.

Area of investigation

Sofia is the 15th largest city in the European Union with a population of about 2 million. It is located in the central part of western Bulgaria. The city is situated between 500 and 700 m above sea level and occupies an area of about 500 km². The climate of Sofia is temperate continental with an average annual temperature of 10.6 °C. The problem of air pollution in the town strongly depends on its geographical location: Sofia valley is tightly enclosed by mountains which reduce the possibility of self-cleaning of the atmosphere.

Monitoring of the parameters

The parameters used in the present investigation are: (i) air pollutants: nitrogen oxide, NO ($\mu g.m^{-3}$); nitrogen dioxide, NO_2 ($\mu g.m^{-3}$); carbon monoxide, CO ($mg.m^{-3}$) and (ii) meteorological indicators: ambient temperature, T (°C); air humidity, W (%); wind velocity, V ($m.s^{-1}$); sun radiation, R ($W.m^{-2}$). The survey was executed in the period April–June during 2017 year. The data were downloaded from the WEB of Sofia Municipality [7]. Location of the seven monitoring stations is shown on Fig. 1.

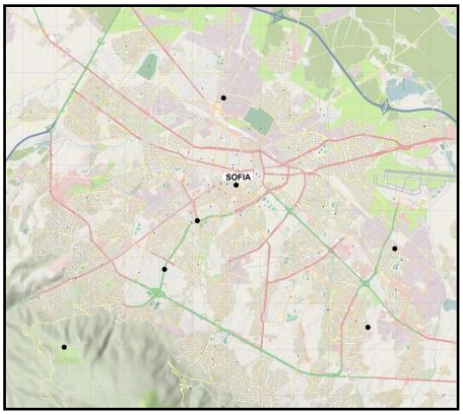


Fig. 1. Location of the automated stations in Sofia, whose data were used in the survey

Modeling method

The multiple linear regression is one of the most commonly used statistical methods for determination of a relationship between air pollution and meteorological parameters. The least squares method were used for calculation of the regression coefficients.

The criteria for the multi collinearity are the values of the regression coefficients in the correlation matrix. When multi collinearity exists, the estimated regression coefficients are not effective (the significance degree of the coefficients is larger than α (0.05). In this case one has to search for some other options to eliminate the negative influence, such as for example, the step-by-step multiple linear regression (MLR) or multiple regression involving new variables of interaction and quadratic functions (MLR + NP). The data were processed by IBM SPSS Statistics 19 and Multi Tab software packages.

The follow indices for validation of the models were used: mean absolute error (MAE), normalized absolute error (NAE), index of agreement (IA), prediction accuracy (PA), coefficient of determination (r^2), mean square error (RMSE) [8].

Computational models

The relations between the concentrations of nitrogen oxides, weather factors and the concentration of CO were modelled by the method of MLR + NP approach, which comprises stepwise multiple regression with quadratic functions and interaction of different variables. The data were modelled without elimination of the large deviations. The models were described by the following equations (Table 1).

Table 1. Equations of the models

Month	***Model***
April	**NO**= – 8.414+ 72.689CO^2 +0.025R– 1.610T*CO- 0.00002059R^2+0.481T
	NO_2= -10.465 + 164.773CO – 67.668CO^2 – 0.399W – 0.061CO*R + 0.019T*W – 1.728T +2.146CO*T + 0.021R – 0.471V^2
	NO_x= –15.634+166.106CO – 0.474W + 0.026T*W-1.295T
May	**NO** = 6.165+67.661CO^2+0.011R-2.061CO*T- -0.000007483R^2+0.039CO*R- 24.711CO-0,001 T*R+0.030T^2
	NO_2= 176.3+ 220.0 CO – 11.01 T – 3.656 W – 14.01 V + 0.0067 R+ + 0.01782 W^2 – 1.781 V^2+ + 4.74 CO*T – 1.466 X1*X3- – 41.66 CO*W – 0.0662 CO*R + 0.07962 T*W + 1.509 T*V+ + 0.2383 W*V + 0.000394 W*R
	NO_x= 108.5+ 236.3 CO- 7.712 T- 2.942 W+ 14.48 V– 0.04565 R+ 83.1 CO^2 + 0.01693 W^2- 3.481 V^2 – 1.820 CO*W – 39.7 CO*V+ + 0.0693 T*W + +1.085 T*V + 0.000785 W*R
June	**NO** = 52.17- 196.9 CO- 0.323 T- 0.3050 W- 0.00373 R+ 159.9 CO^2- -0.000011 R^2+ 0.813 CO*T+ + 0.860 CO*W + 0.02311 CO*R+ + 0.000131 W*R
	NO_2= –175.2 + 75.9 CO + 9.69 T + 1.406 W – 0.56 V + 0.02865 R+ + 123.2 CO^2 – 0.1657 T^2+ 2.98 CO*T – 1.164 CO*W -– 16.94 CO*V – 0.0827 CO*R – 0.0498 T*W+ 0.0942 W*V
	NO_x= –126.9– 185.8 CO+11.21 T+ 1.223 W – 0.0474 R+ + 2.17 V+ 320.7 CO^2- 0.1857 T^2– 0.947 V^2+ 2.93 CO*T- – 0.0537 CO*R – 0.0622 T*W + 0.000838 T*R+ + 0,000798 W*R

The statistical evaluation of the observed and calculated values in the models of nitrogen oxides, for the three months is presented in Table 2.

Table 2. Statistical evaluation of the observed and calculated values

Performance indicator	***April*** (n = 361)			***May*** (n = 491)			***June*** (n = 439)		
	NO	NO_2	NO_x	NO	NO_2	NO_x	NO	NO_2	NO_x
NAE	0.3735	0.1896	0.1893	0.3058	0.2580	0.2505	0.2878	0.2413	0.2355
MAE	3.3325	5.5467	7.2276	1.4295	5.8963	6.8977	1.3614	4.9244	6.0006
RMSE	5.7876	7.9154	11.0481	1.8136	7.2717	8.4195	1.9007	2.3428	7.4514
R^2	0.8527	0.8208	0.8671	0.8139	0.7287	0.7484	0.7683	0.7985	0.8070
IA	0.9591	0.9491	0.9635	0.9067	0.8498	0.8614	0.8815	0.8921	0.8966
PA	0.8564	0.8231	0.8681	0.8816	0.7299	0.7486	0.7570	0.7990	0.8068
Range of VIF	1.0153 –3.112	1.0109 –7.830	1.0173 –2.7777	1.0215 –7.6103	1.0215 –5.1177	1.0146 –5.4171	1.0270 –2.815	1.0221 –3.6483	1.0319 –7.834

n – timepoints number

The results revealed that all models have no problem with multicollinearity, the values of the factors VIF (SPSS-test) are lower than ten. The coefficients of determination for the various models range from 0.73 to 0.87, i.e. 73% to 87% of the nitrogen oxides content changes can be explained by the factors T; W; V; R; CO and their quadratic functions and interactions.

The best fitted models for assessment of the nitrogen oxides pollution are for April. They have the highest levels of accuracy: IA from 0.9491 to 0.9635; PA from 0.8231 to 0.8681; coefficients of determination r^2 from 0.8208 to 0.8671 and the lowest errors: NAE from 0.1893 to 0.3735; MAE from

3.3325 to 7.2276 and RMSE from 5.7876 to 11.0481. For May and June the statistical estimates exhibit somewhat lower accuracy.

The predicted values of the meteorological parameters up to 25 days could be taken from web based platforms. The concentration of CO can be predicted using appropriate models or determined by a process of elimination. For the purpose, any the least probable CO values for the month, time and forecasted meteorological data were consecutively eliminated. The remaining values of CO were averaged. They were assumed as predicted values for CO and were designated as CO_{pr}.

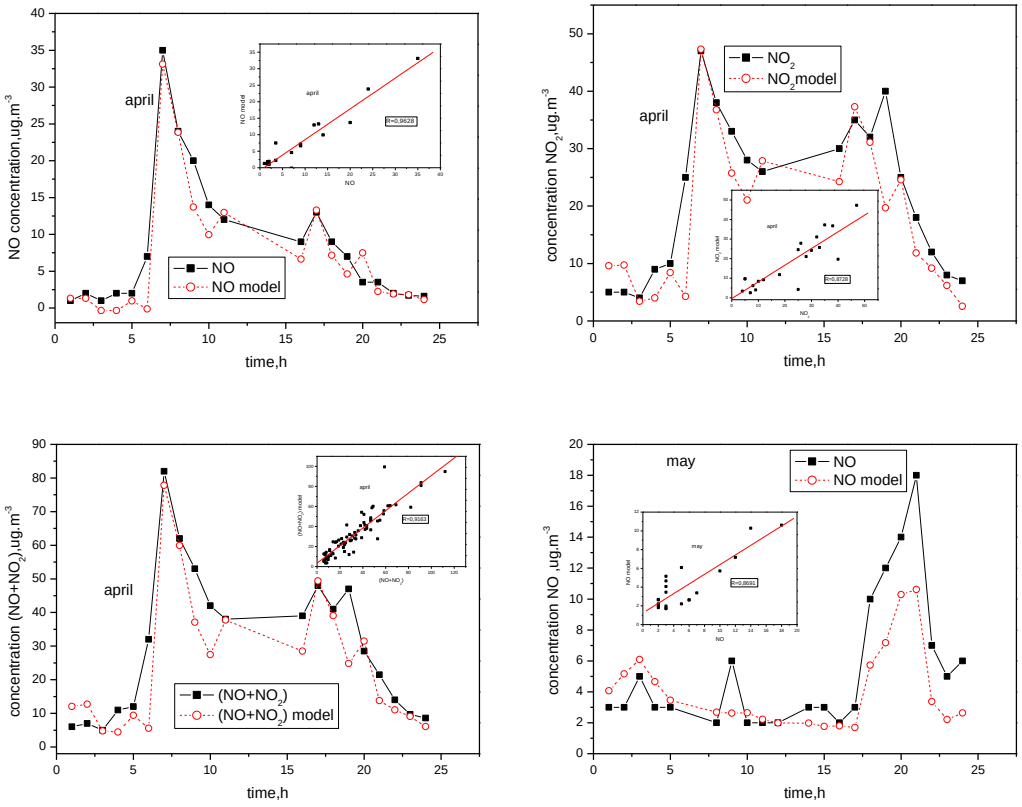
Data from the performed short-term assessment, concerning 24-hour (one day) period were not used in the modeling process. The predicted values for concentrations of nitrogen oxides were calculated using models, based on meteorological forecast indicators and the most likely forecasted concentration of CO, CO_{pr}. The predictions are compared with the measured concentrations of the same oxides and are presented in Fig. 2. The statistical evaluation is given in Table 3. In general, there is a very good quality of the prediction made by the MLR - NP method.

Table 3. Statistical evaluation

Performance indicator	***April*** (n = 20)			***May*** (n = 21)			***June*** (n = 23)		
	NO	NO_2	NO_x	NO	NO_2	NO_x	NO	NO_2	NO_x
NAE	0.2289	0.2282	0.2259	0.3955	0.2179	0.1576	0.2632	0.2681	0.2498
MAE	1.9492	4.9869	6.8608	2.1470	6.3437	4.1051	1.0037	3.8936	4.3282
RMSE	2.7653	7.5083	9.9847	2.8342	5.9206	5.0608	1.2086	5.2195	5.1290
R^2	0.9269	0.7618	0.8385	0.7553	0.7256	0.8077	0.7670	0.7501	0.7600
IA	0.9746	0.9159	0.9403	0.6407	0.7962	0.8920	0.8749	0.8060	0.8576
PA	0.9623	1.0300	0.9472	0.4523	0.9180	0.9633	1.6320	0.9082	0.8694

n – timepoints number

The good assessment of the peak pollution with nitrogen oxides for all months should be noted, and in most models the real hourly values are lower than the predicted ones.



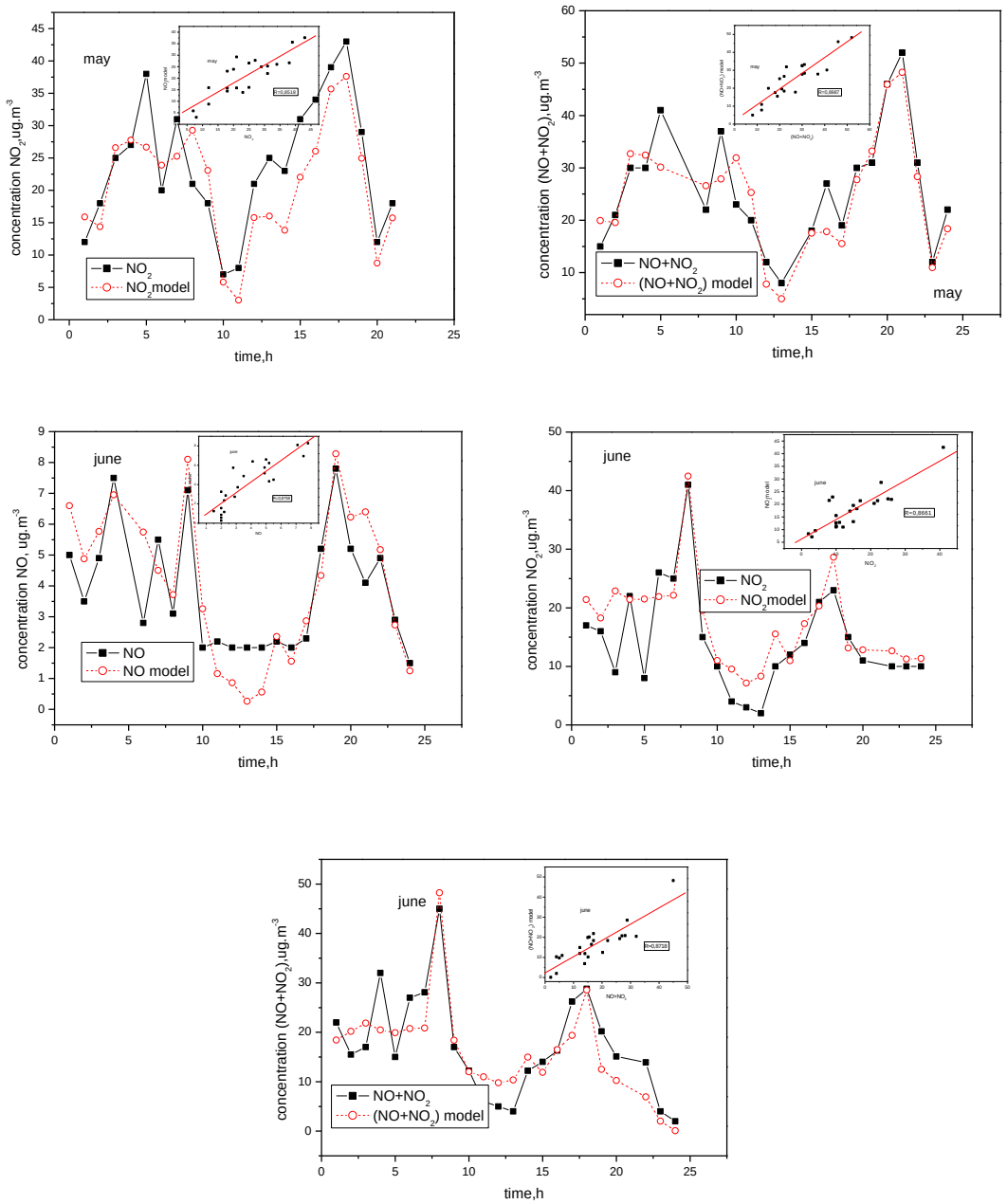


Fig. 2. Comparison between the observed and the predicted concentrations of the nitrogen oxides and their correlation by using MLR (the embedded figures) for April, May, June and for 24 h time duration

The relative errors in forecasting of the maximum concentration of nitrogen oxides are presented in Table 4. The obtained results reveal that the prevailing error values for most models are also below 10 %. The forecasting of the maximum NO concentration for May has the highest relative errors, 22 % and 41 % respectively.

Table. 4. Relative errors for prediction of maximum concentrations of nitrogen oxides over a 24-hour period

Month	***April***		***May***		***June***	
Relative error, %	*time*	*time*	*time*	*time*	*time*	*time*
	7 a.m.	5 p.m.	11 a.m.	6 p.m.	9 a.m.	11 p.m.
NO	5.7	–1.5	41.1	–22	–6.4	–14.8
NO_2	–0.6	–6.6	12.6	–6.4	–24.3	–3.6
NO_x	5.1	–2.9	7.3	–9.0	1.0	–7.1

Conclusions

A statistical investigation was carried out over a three months period deals with the air pollution with nitrogen oxides in Sofia. Empirical models which possess very good adequacy were obtained using the MLR - NP method. In the templates besides linear and quadratic functions, their interactions also were included. The models are characterized by high accuracy. The best fitted models for predicting the air pollution with nitrogen oxides are those for April.

The reliability of the designed models was tested for prediction of the NO_x pollution. The forecasted values for time (t_{pr}); temperature (T_{pr}); wind speed (V_{pr}); humidity (W_{pr}) and radiation (R_{pr}) were used as variables in the test. The estimated concentration of CO (CO_{pr}) used in the models was also determined on the basis of the assessment of the mentioned meteorological indicators.

It was established that most of the models predict the maximum concentration of nitrogen oxides with an error below 10 %. The models proposed in this study predict the concentrations of nitrogen oxides based on only four meteorological indicators which are available in weather forecast web sites. They could be successfully used for prediction and timely information about air pollution with NO_x, both for short and long periods of time. The method MLR-NP was applied is suitable for drawing short-term forecasting models about the content of nitrogen oxides and provides possibilities of prevention of air pollution in large settlements.

References:

1. Gvozdic, V., E. Kovac-Andric, J. Brana. Influence of meteorological factors NO_2, SO_2, CO and PM_{10} on the concentration of O_3 in the urban atmosphere of Eastern Croatia, Environ. Model. Assess. 16, 2011, pp. 491–501.
2. Kumar, A., P. Goyai. Forecasting of air quality in Delhi using principal component regression technique, Atmos. Pollut. Res. 2, 2011, pp.436–444.
3. Slini, T., K. Karatzas, N. Moussiopoulos, Correl ation of air pollution and meteorological data using neural networks" ,8-th Int. Conf. on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, 1998, pp. 368–372.
4. Statheropoulos, M., N. Vassiliadis, A. Pappa. Principal component and canonicalcorrelation analysis for examining air pollution and meteorological data, Atmospheric environmentvol. 32, 6, 1998, pp. 1087–1095.
5. Kalbarczyk, R., E. Kalbarczyk. Influence of meteorological conditions on the concentration of NO_2 and NO_x in northwest Poland in relation to wind direction. Annals of Warsaw University of Life Sciences – SGGW, Land Reclamation, 38, 2007, pp.81–94.
6. Brunelli, U., V. Piazza, L. Pignato, F. Sorbello, S. Vitabile. Two-days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo Italy, Atmospheric Environment, 41, 14, 2007, pp. 2967–2995.
7. Cai, M., Y. Yin, M. Xie. Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach,Transport Res. D-TR E 14, 2009, pp. 32–41.
8. On line at: https://www.sofia.bg/components-environment-air
9. Lu, H. Estimating the emission source reduction of PM_{10} in Central Taiwan, Chemosphere 54, 2004, pp. 805–814.