Inżynieria i Ochrona Środowiska 2018, 21(4), 321-333
Engineering and Protection of Environment
https://ios.is.pcz.pl/

p-ISSN 1505-3695
e-ISSN 2391-7253
DOI: 10.17512/ios.2018.4.1

Szymon HOFFMAN, Mariusz FILAK

Czestochowa University of Technology, Faculty of Infrastructure and Environment
Department of Chemistry, Water and Wastewater Technology
ul. J.H. Dąbrowskiego 73, 42-201 Częstochowa
e-mail: szymon@is.pcz.czest.pl

Prediction of Monthly Averages of Air Pollutant Concentrations for Selected Areas in Mazovian Voivodeship

Predykcja średniomiesięcznych stężeń zanieczyszczeń powietrza dla wybranych obszarów województwa mazowieckiego

The study was carried out using long-term data, recorded at two air monitoring stations in Masovian Voivodeship. Hourly time series, obtained from the monitoring system, were averaged in calendar months to get monthly time series. The data sets, containing time series of monthly mean values from two different monitoring sites, were subjected to multivariate regression analysis. Models of multidimensional linear regression were built for the both sets of data. The obtained models describe statistical dependencies between concentrations of specified air pollutants and concentrations of other pollutants and meteorological parameters, recorded at the same monitoring station. The achieved regression equations were used to predict long-term courses of monthly concentrations. For visualization of prediction accuracy, the charts containing time series of actual and predicted monthly concentrations were prepared. The approximation precision was estimated by calculating modelling errors for each regression model. Three different measures of approximation error were applied: mean absolute error (MAE), root mean square error (RMSE), and Pearson correlation coefficient (r).

Keywords: **air pollution, air monitoring, pollutant concentrations, monthly concentrations, multivariate regression models, approximation error**

Introduction

The concentration of air pollutants depends on many factors. The most important of them are: local emissions of pollutants, inflow of pollutants from other areas, chemical and physicochemical transformations of pollutants in the air, meteorological conditions, topography of the area. Many factors remain unknown. Others have been identified, but it is difficult to quantify them. Therefore, prediction of the concentration of pollutants is usually fraught with relatively high errors. Despite conceptual differences, prediction models have one thing in common: they explore available data.

Statistical methods are commonly used to determine the complex effect of various factors on the concentration levels of air pollutants [1, 2]. However, statistical

Open Access (CC BY-NC 4.0)

analysis imposes a specific structure for all explored variables. It is advisable that the data to be analysed should be as uniformly structured as possible and come from the same source. This criterion is met by data recorded at automatic air quality monitoring stations. In these facilities, all variables, both concerning concentrations and meteorological conditions, are measured continuously and then recorded as time series of 1-hour mean values. The data with such a structure can be explored by means of two main methods: regression methods and time series methods [3]. The aim of exploration may be looking for statistical relationships hidden in data sets. The identified dependencies can be used in practice, for example to fill in missing data, i.e. data not recorded in monitoring systems [3]. Data collected at air monitoring stations are incomplete [4]. The assessment of air quality in a selected area may be difficult or even impossible because of large data gaps.

Multivariate regression models allow for determination of the approximate contribution of each of the independent factors in changes in the concentration of the selected pollutant. On the one hand, a good model should be complete enough to describe a complex phenomenon, and, on the other hand, it should be as simple as possible to be more comprehensible [5]. That is why the linear models are the most commonly used models of multivariate regression. They allow to estimate the influence of individual explanatory variables on the explained variable [6].

The aim of the analysis was to find regression equations for prediction of average monthly concentrations of basic air pollutants at two different air monitoring stations. Although monthly means of concentrations are not used for formal evaluation of air quality, their profiles facilitate visualization of changes in concentrations over long-term measurement periods and can be useful in identifying periods of the greatest environmental risk.

The study was conducted using the long-term measurement data, recorded at two air monitoring stations in Mazovian Voivodeship. One-hour time series were averaged to mean monthly concentrations. In order to approximate concentrations, some regression equations were found, which were used to predict the mean monthly concentrations of air pollutants. Approximated concentrations were compared to actual concentrations in order to evaluate prediction accuracy.

1. Methods

The analysis was made on the data obtained from Voivodeship Inspectorate for Environmental Protection in Warsaw. They come from measurements that have been made over the years in Masovian Voivodeship, under the state program of air monitoring. Two automatic monitoring stations were chosen: the first one in Radom, representing the urban zone of the city of Radom, and the second one in Granica, a village located in the Kampinos National Park, treated as a background station in the voivodeship. The data from the station in Radom covered the period 2005-2016, whereas the data from the station in Granica concerned the period 2004-2016.

The data from both monitoring stations included time series of hourly concentrations of basic air pollutants and time series of hourly values of measured meteorological parameters. Hourly time series were averaged in calendar months to give monthly time series. The data sets containing monthly time series were subjected to multivariate regression analysis. Multivariate linear regression models describing statistical dependence of concentrations of individual air pollutants on concentrations of other pollutants and on meteorological parameters were developed for both stations.

For each monthly concentration time series in both sets, the multivariate regression equation was found in its general form of:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \ldots$$

where:
Y - explained variable,
X_1, X_2, X_3, … - explanatory variables,
β_0, β_2, β_2, β_3, … - sought regression coefficients.

The variables were denoted as follows:

O_3	- monthly mean O_3 concentration, $\mu g/m^3$
NO	- monthly mean NO concentration, $\mu g/m^3$
NO_2	- monthly mean NO_2 concentration, $\mu g/m^3$
SO_2	- monthly mean SO_2 concentration, $\mu g/m^3$
CO	- monthly mean CO concentration, mg/m^3
PM10	- monthly mean concentration of PM10, $\mu g/m^3$
T	- monthly mean temperature, °C
P	- monthly mean intensity of solar radiation, W/m^2
V	- monthly mean wind speed, m/s
W	- monthly mean relative humidity, %

CO and PM10 concentrations were not measured at the monitoring station in Granica. Therefore, regression models with the use of these variables could not be created. Independent variables whose impact on the explained variable was statistically insignificant, were rejected from the models. The significance of independent variables was evaluated by analysing the p-value. Variables whose p-value was less than 0.05 were considered significant. The calculations were made using a Microsoft Excel spreadsheet.

The obtained regression equations were used for predicting monthly concentrations of air pollutants. For visualization of prediction accuracy, the graphs containing time series of observed and predicted monthly concentrations were prepared. The accuracy of approximation was estimated by calculating modelling errors for each regression model. Three different measures of approximation error were applied: mean absolute error (MAE), root mean squared error (RMSE) and the Pearson correlation coefficient (r).

2. Results and discussion

The modelling results are presented in the following subsections, separately for each of the pollutants.

2.1. Modelling of O_3 concentrations

Characteristics of the obtained models of multivariate regression of O_3 concentrations are presented in Table 1. Both regression models have the same explanatory variables (NO, T, P, W). An additional explanatory variable of SO_2 is included in the model M1 from Granica.

Table 1. **Characteristics of multivariate regression models obtained for O_3 concentrations**

Station	Model symbol	Regression equation	Explanatory variables
Granica	M1	$O_3 = 141.7 - 9.99 \cdot NO - 0.56 \cdot T + 0.04 \cdot P - 1.06 \cdot W + 0.59 \cdot SO_2$	NO, T, P, W, SO_2
Radom	M2	$O_3 = 67.33 - 0.81 \cdot NO - 0.77 \cdot T + 0.17 \cdot P - 0.42 \cdot W$	NO, T, P, W

The graphs of observed and predicted monthly ozone concentrations for monitoring stations in Granica and Radom are showed in Figures 1 and 2, respectively. The interruptions in time series occur during periods of missing data for the explained variable or the explanatory variables entering the models. The shapes of both graphs indicate that the regression models, despite their inaccuracy, reflect quite well the seasonal variability of monthly averages of ozone concentration.

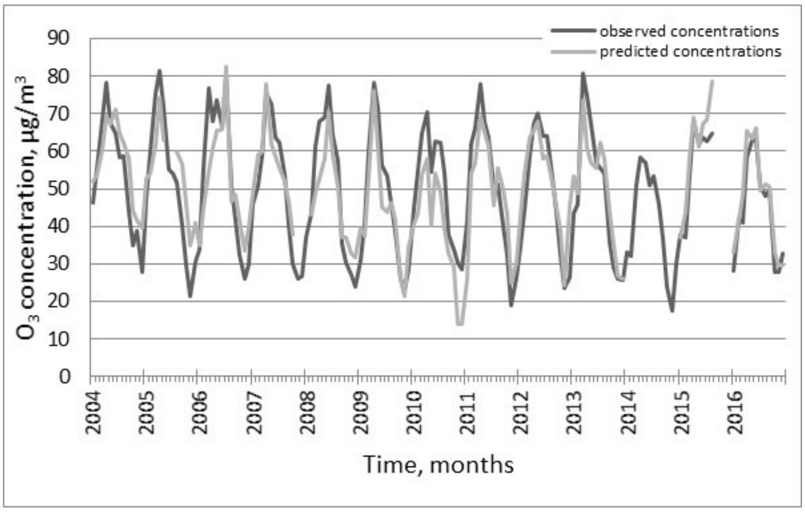

Fig. 1. Graphs of observed and predicted monthly O_3 concentrations for the station in Granica

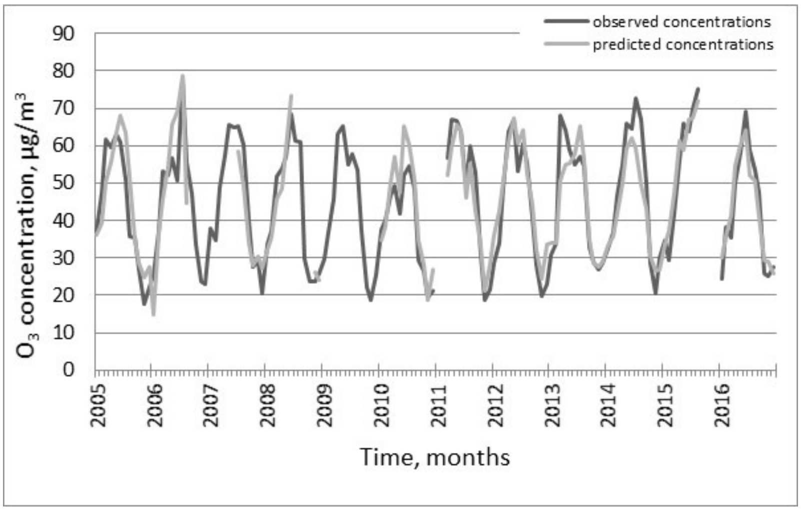

Fig. 2. Graphs of observed and predicted monthly O_3 concentrations for the station in Radom

The values of prediction errors for the regression models of O_3 concentration are presented in Table 2. These values indicate that ozone concentration can be modelled with higher accuracy at the station in Radom (lower MAE and RMSE errors, higher correlation coefficients).

Table 2. **Prediction errors for regression models of O_3 concentration**

Station	Model symbol	Explained variable	MAE µg/m^3	RMSE µg/m^3	r
Granica	M1	O_3	6.7	8.1	0.872
Radom	M2	O_3	5.0	6.3	0.918

2.2. Modelling of NO concentrations

Characteristics of multivariate regression models obtained for NO concentrations are presented in Table 3. Both regression models have the same explanatory variables (O_3, NO_2). An additional explanatory variable V is included in the model M3 from Granica, whereas in the model M4 from Radom the additional variables are T and PM10.

Table 3. **Characteristics of multivariate regression models obtained for NO concentrations**

Station	Model symbol	Regression equation	Explanatory variables
Granica	M3	$NO = 1.55 - 0.01 \cdot O_3 + 0.08 \cdot NO_2 - 0.32 \cdot V$	O_3, NO_2, V
Radom	M4	$NO = -0.17 - 0.16 \cdot O_3 + 0.28 \cdot NO_2 + 0.15 \cdot PM10 + 0.26 \cdot T$	O_3, NO_2, PM10, T

The graphs of observed and predicted monthly NO concentrations for monitoring stations in Granica and Radom, are showed in Figures 3 and 4, respectively. The discontinuities in time series appear during periods of missing data for an the explained variable or the explanatory variables entering the models. Comparison of the profiles of observed and predicted concentrations reveals that the regression models significantly reduce predicted concentrations for the months in which observed concentrations are high. The shapes of the graphs indicate that, despite substantial inaccuracy, the regression models reflect the seasonal variability of monthly averages of NO concentration.

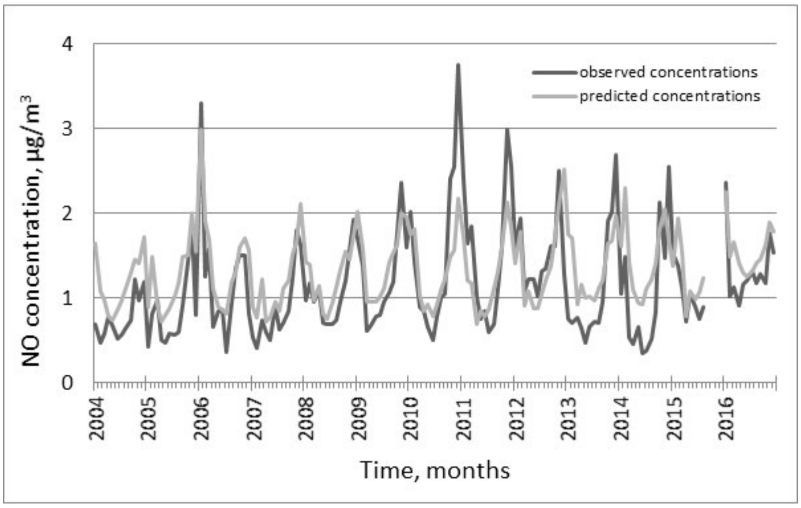

Fig. 3. Graphs of observed and predicted monthly NO concentrations for the station in Granica

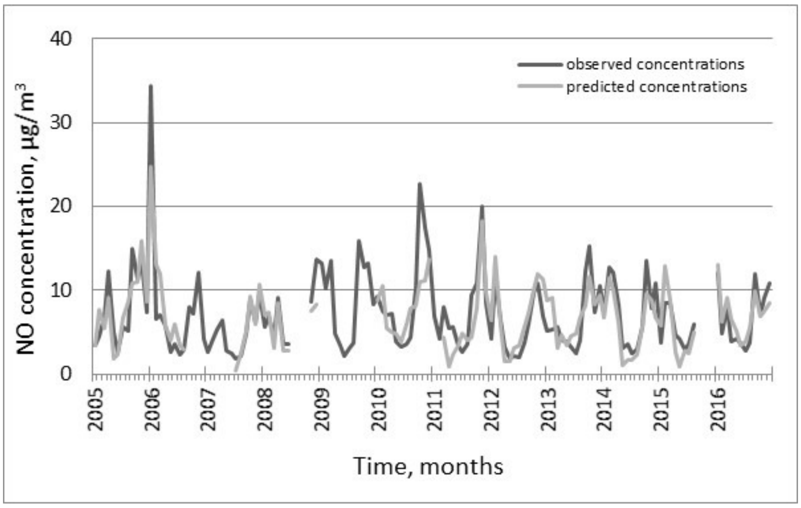

Fig. 4. Graphs of observed and predicted monthly NO concentrations for the station in Radom

The values of prediction errors for the regression models of NO concentration are presented in Table 4. MAE and RMSE errors are higher in the case of the M4 model from Radom, but the level of concentrations recorded at this station is also many times higher than that in the station in Granica. The value r demonstrates that the M4 model is more accurate than the M3 model.

Table 4. **Prediction errors for regression models for NO concentration**

Station	Model symbol	Explained variable	MAE µg/m^3	RMSE µg/m^3	r
Granica	M3	NO	0.4	0.4	0.753
Radom	M4	NO	2.0	2.7	0.826

2.3. Modelling of NO_2 concentrations

Characteristics of multivariate regression models obtained for NO_2 concentrations are presented in Table 5. Only one explanatory variable is common to both models M5 and M6 (that is NO concentration). NO concentration is usually positively correlated with NO_2 concentration. It can be explained by the origin of both gases from the same sources of pollution, i.e. combustion processes.

Table 5. **Characteristics of multivariate regression models obtained for NO_2 concentrations**

Station	Model symbol	Regression equation	Explanatory variables
Granica	M5	$NO_2 = 12.82 + 1.71 \cdot NO + 0.2 \cdot SO_2 - 3.14 \cdot V - 0.32 \cdot T$	NO, SO_2, V, T
Radom	M6	$NO_2 = 12.11 + 0.28 \cdot NO + 14.65 \cdot CO$	NO, CO

The graphs of observed and predicted monthly NO_2 concentrations for monitoring stations in Granica and Radom are illustrated in Figures 5 and 6, respectively. The discontinuities in time series occur during periods of missing data for the explained variable or the explanatory variables. The shapes of both graphs indicate that the regression models reflect the seasonal variability of monthly mean NO_2 concentrations quite well in the case of monitoring stations in Granica (Fig. 5) and much worse in the case of Radom (Fig. 6).

The values of predicted errors for the regression models of NO_2 concentration are presented in Table 6. All error categories indicate that the modelling of NO_2 concentrations for the station in Radom is burdened with a high error.

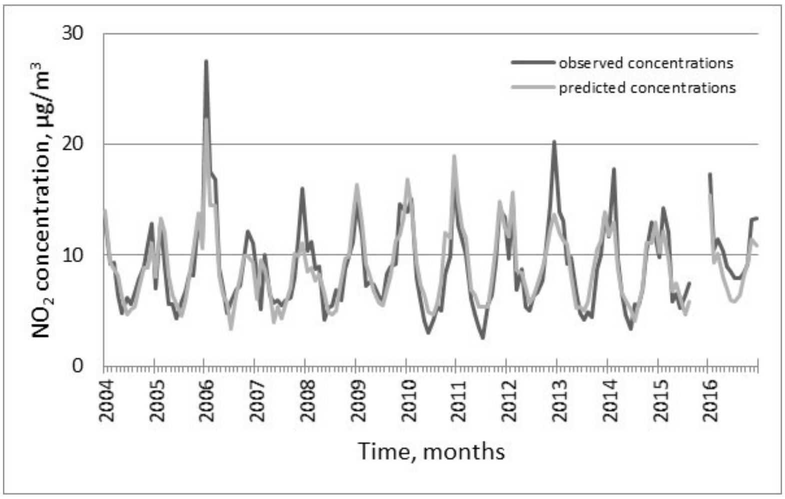

Fig. 5. Graphs of observed and predicted monthly NO_2 concentrations for the station in Granica

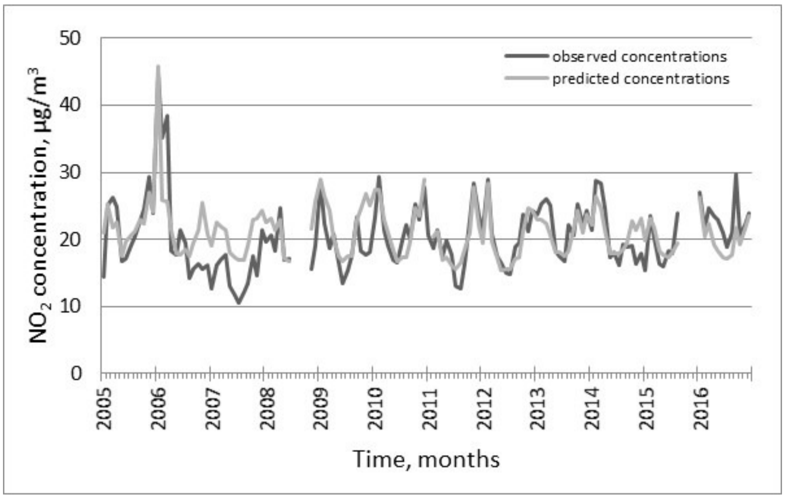

Fig. 6. Graphs of observed and predicted monthly NO_2 concentrations for the station in Radom

Table 6. **Prediction errors for regression models of NO_2 concentration**

Station	Model symbol	Explained variable	MAE μg/m^3	RMSE μg/m^3	r
Granica	M5	NO_2	1.4	1.8	0.886
Radom	M6	NO_2	2.9	3.9	0.627

2.4. Modelling of SO_2 concentrations

Characteristics of multivariate regression models obtained for SO_2 concentrations are presented in Table 7. Only one explanatory variable is common for both models M7 and M8 (that is O_3 concentration).

Table 7. **Characteristics of multivariate regression models obtained for SO_2 concentrations**

Station	Model symbol	Regression equation	Explanatory variables
Granica	M7	$SO_2 = -2.54 + 0.04 \cdot O_3 + 0.73 \cdot NO_2$	O_3, NO_2
Radom	M8	$SO_2 = -8.30 + 0.04 \cdot O_3 + 22.9 \cdot CO + 0.08 \cdot PM10$	O_3, CO, PM10

The graphs of observed and predicted monthly SO_2 concentrations for monitoring stations in Granica and Radom are presented in Figures 7 and 8, respectively. The discontinuities in time series occur during periods of missing data for the explained variable or the explanatory variables. The shapes of the graphs indicate that the regression models reflect the characteristic seasonal variability of SO_2 monthly concentration. Overestimation of predicted concentrations can be visible in the last few years, but underestimation of these concentrations occur in the years at the beginning of the graphs. Furthermore, the gradual decline in the concentration of this pollutant is also observed.

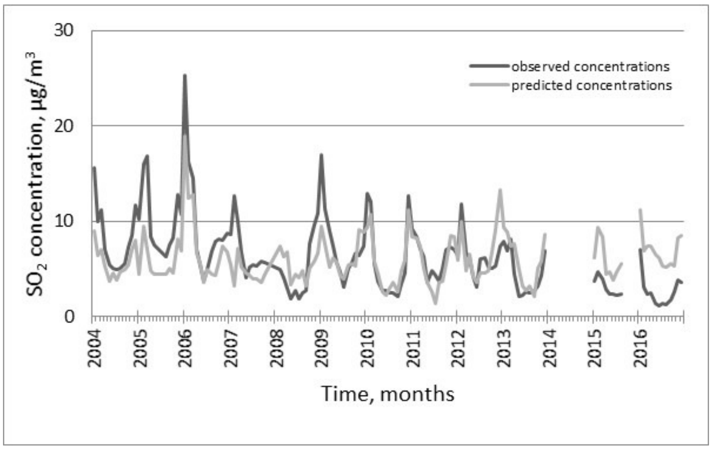

Fig. 7. Graphs of observed and predicted monthly SO_2 concentrations for the station in Granica

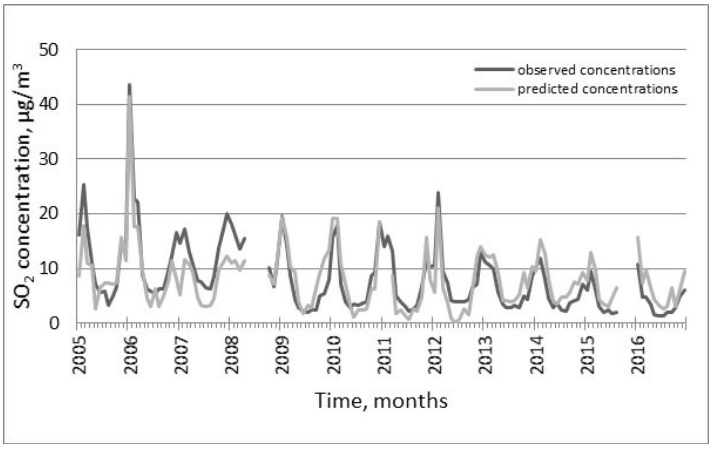

Fig. 8. Graphs of observed and predicted monthly SO_2 concentrations for the station in Radom

The values of prediction errors for the regression models of SO_2 concentration are presented in Table 8. The values of MAE and RMSE errors are higher in the case of the M8 model from Radom, but the level of concentrations recorded at this station is also higher than that in the station in Granica. The value r demonstrates that the M8 model is more accurate than M7.

Table 8. **Prediction errors for regression models of SO_2 concentration**

Station	Model symbol	Explained variable	MAE $\mu g/m^3$	RMSE $\mu g/m^3$	r
Granica	M7	SO_2	2.3	3.0	0.652
Radom	M8	SO_2	2.7	3.3	0.849

2.5. Modelling of CO concentrations

The CO concentration model was created only for the data from Radom. At the station in Granica, CO measurement was not included in monitoring programs. Characteristics of multivariate regression models obtained for CO concentrations are presented in Table 9. The M9 model has three explanatory variables (NO_2, SO_2, PM10). All regression factors are positive, which means that monthly CO concentrations are positively correlated with the concentrations of the mentioned pollutants.

Table 9. **Characteristics of multivariate regression models obtained for CO concentrations**

Station	Model symbol	Regression equation	Explanatory variables
Radom	M9	$CO = 0.02 + 0.007 \cdot NO_2 + 0.012 \cdot SO_2 + 0.005 \cdot PM10$	NO_2, SO_2, PM10

The graphs of observed and predicted monthly CO concentrations for the monitoring station in Radom are showed in Figures 9. The discontinuities in time series occur during periods of missing data for the explained variable or the explanatory variables. The shapes of the profiles presented in the graph indicate that the model M9 reflects the seasonal variability of monthly CO concentration very well.

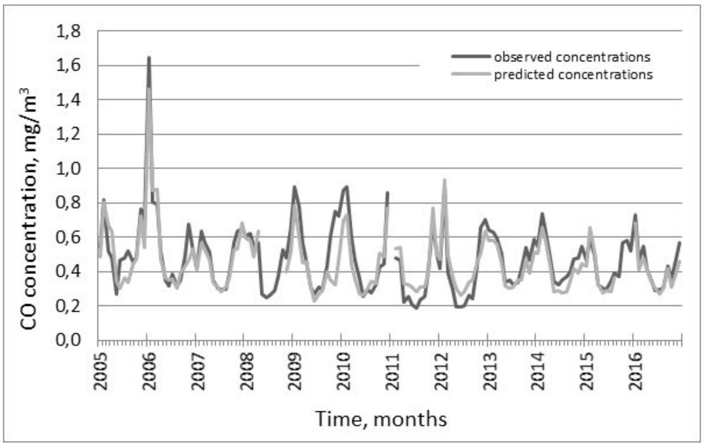

Fig. 9. Graphs of observed and predicted monthly CO concentrations for the station in Radom

The values of prediction errors for the regression models of CO concentration are presented in Table 10. Values of modelling accuracy measures indicate that the monthly CO concentration at the station in Radom can be modelled with high accuracy.

Table 10. **Prediction errors for regression models for CO concentration**

Station	Model symbol	Explained variable	MAE µg/m^3	RMSE µg/m^3	r
Radom	M9	CO	0.1	0.1	0.902

2.6. Modelling of PM10 concentrations

The PM10 concentration model was created only for the data from Radom. At the station in Granica, PM10 measurement was not included in monitoring programs. Characteristics of the model of multivariate regression of PM10 concentrations are presented in Table 11. The model M10 has five explanatory variables (NO, CO, SO_2, V, T).

Table 11. **Characteristics of multivariate regression models obtained for PM10 concentrations**

Station	Model symbol	Regression equation	Explanatory variables
Radom	M10	$PM10 = 23.3 + 0.78 \cdot NO + 35.2 \cdot CO + 0.50 \cdot SO_2 - 5.08 \cdot V - 0.47 \cdot T$	NO, CO, SO_2, V, T

The graphs of observed and predicted monthly PM10 concentrations for the monitoring station in Radom are illustrated in Figure 10. The discontinuities in time series occur during periods of missing data for the explained variable or the explanatory variables. The shapes of the profiles presented in the graph demonstrate that the regression model M10 reflects the seasonal variability of monthly PM10 concentration very well.

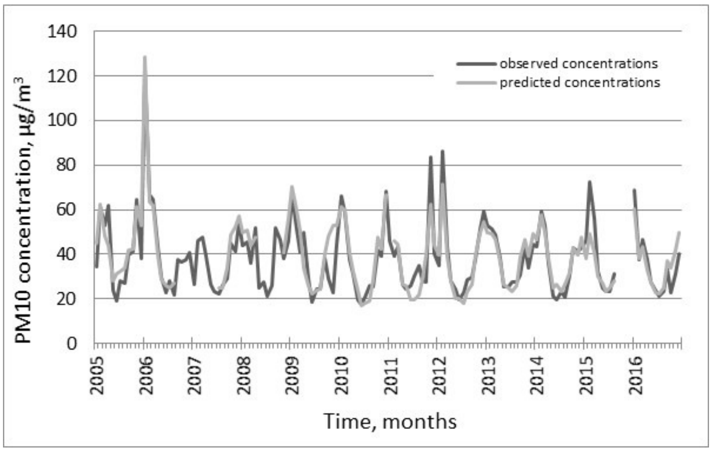

Fig. 10. Graphs of observed and predicted monthly PM10 concentrations for the station in Radom

The values of prediction errors for the regression model of PM10 concentration are presented in Table 12. The values of modelling accuracy measures indicate that monthly PM10 concentration at the station in Radom can be modelled with high accuracy. MAE and RMSE values are high, but overall PM10 concentrations are also high.

Table 12. **Prediction errors for regression models for PM10 concentration**

Station	Model symbol	Explained variable	MAE µg/m^3	RMSE µg/m^3	r
Radom	M10	PM10	5.1	7.2	0.901

Conclusions

The following most important conclusions can be drawn from the analysis presented in the study:

1. For each of the basic air pollutants, measured at the air monitoring station, a statistically significant multivariate regression model can be found. Such a model allows prediction of pollutant concentration by means of explanatory variables that are recorded at the same air monitoring station. Predictive variables for the selected pollutant are specific for the monitoring stations and do not have to be repeated in other sites.
2. Modelling errors differ for various pollutants. The most accurate models can be found for monthly mean concentrations of O_3, CO and PM10 (models: M2, M9, M10). The least accurate models are M6 and M7, for NO_2 in Radom and SO_2 in Granica, respectively.
3. All regression models reflect the seasonal variation in the courses of monthly mean concentrations.

Acknowledgments

This study was financed by Czestochowa University of Technology within statutory research grant BS-PB-402/-301/15.

References

[1] Milionis A.E., Davies T.D., Regression and stochastic models for air pollution - I. Review, comments and suggestions, Atmos. Environ. 1994, 28, 17, 2801-2810.

[2] Milionis A.E., Davies T.D., Regression and stochastic models for air pollution - II. Application of stochastic models to examine the links between ground-level smoke concentrations and temperature inversions, Atmos. Environ. 1994, 28, 17, 2811-2822.

[3] Hoffman S., Jasiński R., Uzupełnianie brakujących danych w systemach monitoringu powietrza, Wydawnictwo Politechniki Częstochowskiej, Częstochowa 2009.

[4] Hauck H., Kromp-Kolb H., Petz E., Requirements for the completeness of ambient air quality data sets with respect to derived parameters, Atmos. Environ. 1999, 33, 2059-2066.

[5] Birkes D., and Dodge Y., Alternative Methods of Regression, Wiley-Interscience Publication, New York 1993.

[6] Zeliaś A., Pawełek B., Wanat S., Prognozowanie ekonomiczne: teoria, przykłady, zadania, WN PWN, Warszawa 2013.

Streszczenie

Badania przeprowadzono, wykorzystując wieloletnie dane pomiarowe zarejestrowane na dwóch stacjach monitoringu powietrza w województwie mazowieckim. 1-godzinne serie czasowe uśredniono w okresach miesięcznych, uzyskując średniomiesięczne serie czasowe. Zbiory danych zawierających serie czasowe wartości średniomiesięcznych poddano analizie regresji wielowymiarowej. W obu zbiorach szukano modeli wielowymiarowej regresji liniowej, opisujących statystyczną zależność stężeń poszczególnych zanieczyszczeń powietrza od stężeń pozostałych zanieczyszczeń i od parametrów meteorologicznych. Otrzymane równania regresji wykorzystano do predykcji średniomiesięcznych stężeń zanieczyszczeń powietrza. Sporządzono wykresy zawierające serie czasowe rzeczywistych i przewidywanych stężeń średniomiesięcznych, które pozwoliły na wizualizację dokładności predykcji. Oszacowano również dokładność aproksymacji, obliczając błędy modelowania dla każdego z modeli regresyjnych. Zastosowano trzy różne miary błędu aproksymacji, obliczając dla modeli regresyjnych średni błąd bezwzględny (MAE), pierwiastek z błędu średniokwadratowego (RMSE), współczynnik korelacji Pearsona (r).

Słowa kluczowe: **zanieczyszczenia powietrza, monitoring powietrza, stężenia zanieczyszczeń, stężenia średniomiesięczne, modele regresji wielowymiarowej, błąd aproksymacji**