Environment Protection Engineering

Vol. 38 2012 No. 2

DOI: 10.5277/epe120210

SZYMON HOFFMAN*

APPROXIMATION OF IMISSION LEVEL AT AIR MONITORING STATIONS BY MEANS OF AUTONOMOUS NEURAL MODELS

Long-term collection of data, recorded at several air monitoring stations located in Central Poland, was analyzed. The main objective of the analysis was to choose optimum modelling methods for concentration of specified air pollutants. For this purpose accuracies of various groups of autonomous models were compared. Prediction of any air pollutants was performed using three different modelling methods. The modelled value was instantaneous concentration of specified pollutant. The models varied in the number and type of the explanatory variables and the modelling technique. It was presumed that there is a need for modelling the measurement gap, comprising a selected extract of the time series of a chosen pollutant. For successive cases in the gap, prediction errors of various methods of modelling were compared.

1. INTRODUCTION

The data gathered continuously in the air monitoring systems are never entire. A need for imputation of modelled values into measurement gaps appears when the completeness of the analyzed monitoring set is too low [1, 2]. Deficient measurement series should not be used for air quality assessment [3]. Operative law does not recommend specific modelling methods. Linear interpolation is the simplest imputation technique. However, this method is not precise enough for longer gaps in monitoring time series. More accurate could be the methods which exploit the knowledge collected in historical data, at the same or at the neighbouring monitoring stations. Such models, which provide prediction without recourse to any data coming from the out of a monitoring system, were called autonomous models [4]. Two main groups of the autonomous models could be specified – regression models and time series models. The regression models exploit relationships between elements from different measure series, whereas the time series models base on autoregression in specified measure series. A choice of method should ensure possibly highest prediction accuracy.

*Częstochowa University of Technology, Department of Chemistry, Water and Wastewater Technology, ul. Dąbrowskiego 69, 42-200 Częstochowa, Poland; e-mail: szymon@is.pcz.czest.pl

The main objective of the analysis was to choose optimum modelling methods for concentration of specified air pollutants. For this purpose, accuracies of various groups of autonomous models were compared. It was presumed that there is a need for modelling a measurement gap, comprising a selected extract of the time series of a chosen pollutant. Additionally, it was assumed that complete data for other pollutants and meteorological parameters during the period covering the gap as well as data from the modelled time series prior to that gap are available. For successive cases in the gap, prediction errors of different methods of modelling were compared. In regression models, available data for a considered case were exploited, i.e. the data recorded on the same day and at the same hour. In time series models, successive cases in a measuring gap were treated as sequent steps of prognosis. To build predictive models neural networks were used because they allow acquiring optimum approximation of modelled variables.

In the present study, the long-term collection of data, recorded at several air monitoring stations located in Central Poland, was analyzed. Approximation was conducted for concentration levels, recorded at the air monitoring station in Radom.

2. METHODS

The analyzed data sets derived from 8 different air monitoring sites in Central Poland, i.e. from Widzew, Gajew, Granica, Piotrków Trybunalski, Legionowo, Radom, Tłuszcz, Ursynów. The data were gathered in the period 2004–2008. Data collection was obtained from Voivodeship Inspectorates for Environmental Protection in Warsaw and in Łódź. The examined time-series involved hourly concentrations of main air pollutants: O_3, NO, NO_2, PM_{10}, SO_2, CO, as well as hourly averages of the following meteorological data: temperature, wind speed, wind direction, relative humidity, and solar radiation.

Approximation of pollutant concentrations was carried out for the monitoring station Radom. The quality of modelling was estimated by comparing the predicted concentrations with the actual ones. Prediction of any air pollutants was performed using various modelling methods. In all methods, the modelled value (model output) was concentration of specified pollutant at a given time (day and hour). The models varied in the number and type of the explanatory variables and the modelling technique. Three basic groups of models were created:

Time-series linear models (TS-L). Concentrations of chosen pollutant recorded in previous hours were inputs of the TS-L models. All time series models had a fixed number of steps (24). The number of steps was the parameter which specified how many previous cases is taken as the inputs to the model. Within the group of TS-L models, several models differing in the horizon of the forecast (lookahead) were gen-

erated. Lookahead was a parameter which indicated the number of steps between the last known value before the gap and the predicted value in the gap. The following lookaheads were assumed: 1, 2, 3, 4, 5, 6, 8, 12, 24. That range of the prognosis horizon allowed predicting concentrations in 24-hour gaps. Linear neural networks were used as a tool of time series analysis. Previous experience showed that nonlinear models did not improve significantly the quality of modelling [5].

Multiple regression perceptron models (MR-P). In these models, as predictors of specified pollutant concentration concentrations of other pollutants were used measured at the same monitoring stations, as well as date and hour of measurement and meteorological data. Non-linear neural networks of the perceptron structure were applied for regression analysis. Each model used a network with five neurons placed in a single hidden layer. Such a relatively simple neural network structure allows effective exploration of knowledge hidden in data [6].

External Multiple Regression Perceptron Models (EMR-P). In these models, as predictors of specified pollutant concentration were used concentrations of the same pollutant measured at other, neighbouring monitoring stations. For example, to model the concentrations of O_3 at the monitoring station Radom, concentrations of O_3 from seven other stations located in Central Poland (Fig. 1) were used. A neural network structure was similar to the network in MR-P models.

Fig. 1. The location of examined air monitoring sites in Central Poland

The analysis was carried out using the Statistica Data Mining application. In each neural network, the analyzed set of data was divided into three subsets: the training subset (50% of cases), the verification subset (25% of cases) and the test subset (25%

of cases). For neural network training two algorithms were used: Levenberg-Marquardt in the group of non-linear models, and pseudo-inverse in the case of linear models. The values of the model errors were estimated when comparing divergences between the model output and the real concentration values. The following categories of prediction error were taken into account:

- the value of Pearson's correlation coefficient (r),
- the value of the root mean square error (RMSE),
- the value of the mean absolute error ($|e|$),
- the ratio RMSE/s, where s is the standard deviation.

The values of RMSE and $|e|$ were useful criteria for assessment the models generated for a specified pollutant at a specified monitoring site. To compare the models for different pollutants more universal are dimensionless errors, like correlation coefficient or the ratio RMSE/s.

3. RESULTS

The changes of the values of modelling errors in dependence on the prognosis lookahead were shown in figures 2-7. Individual figures illustrate the results obtained successively for O_3, NO, NO_2, CO, SO_2, PM_{10}. Curves of changes in four different categories of error: RMSE, $|e|$, RMSE/s i r were presented for the forecast horizons ranging from 1 to 24 hours. Each chart comprises the results for the models generated by 3 different prediction techniques, including models of time series (TS-L), non-linear multivariate regression models (MR-P), and non-linear multivariate regression models exploring data from adjacent monitoring stations (EMR -P). Exact values of the prediction errors were given in an earlier publication [7].

4. DISCUSSION

Regression models MR-P and EMR-P have constant errors, independent of the forecast horizon. Only the TS-L models are characterized by a variable value of prediction errors in a measuring gap. For those models, the RMSE, $|e|$, RMSE/s errors gradually increase with increasing forecast horizon. Correlation coefficient values behave different; their decrease means prediction quality drops. Analysis of any measurement errors leads to the conclusion that the quality of modelling in the group of models TS-L strongly depends on the length of the measuring gap. That regularity is observed for all air pollutants. Only for short gaps TS-L model can be more accurate than others. Tables 1–6 summarize separately for individual air pollutants which modelling method is the most accurate for measuring gaps of various lengths. Obtained results can be just as well referred to subsequent cases of a 24-hour long gap.

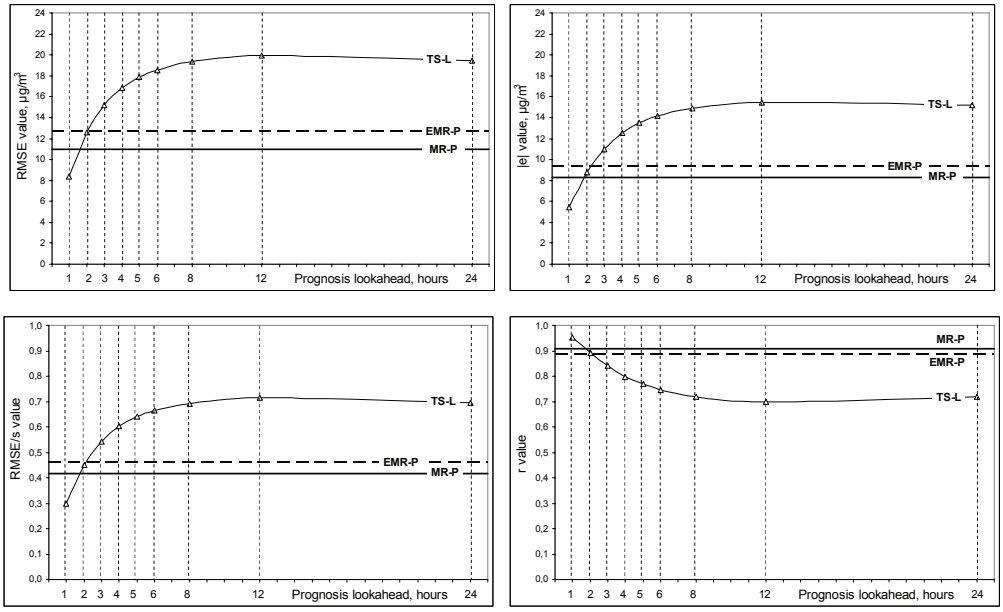

Fig. 2. Modelling errors of O_3 concentrations in dependence on the forecast horizon (Models TS-L, MR-P, EMR-P, Radom 2004–2008)

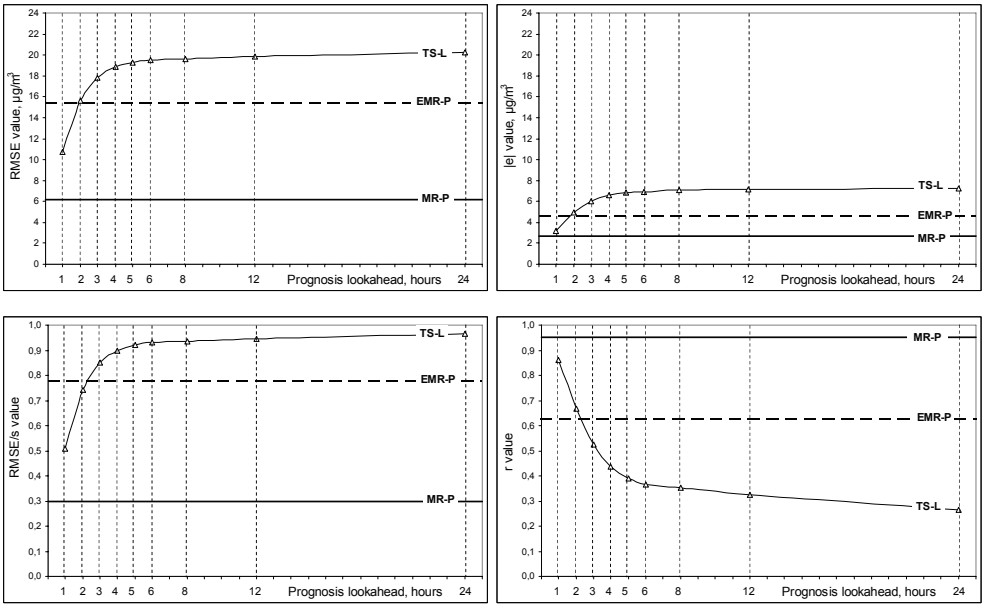

Fig. 3. Modelling errors of NO concentrations in dependence on the forecast horizon (Models TS-L, MR-P, EMR-P, Radom 2004–2008)

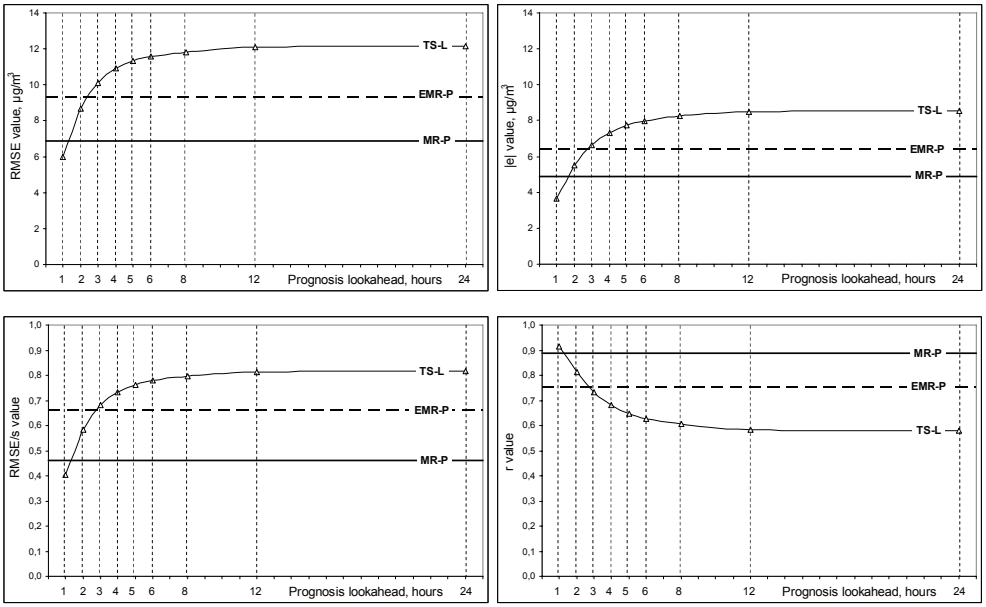

Fig. 4. Modelling errors of NO_2 concentrations in dependence on the forecast horizon (Models TS-L, MR-P, EMR-P, Radom 2004–2008)

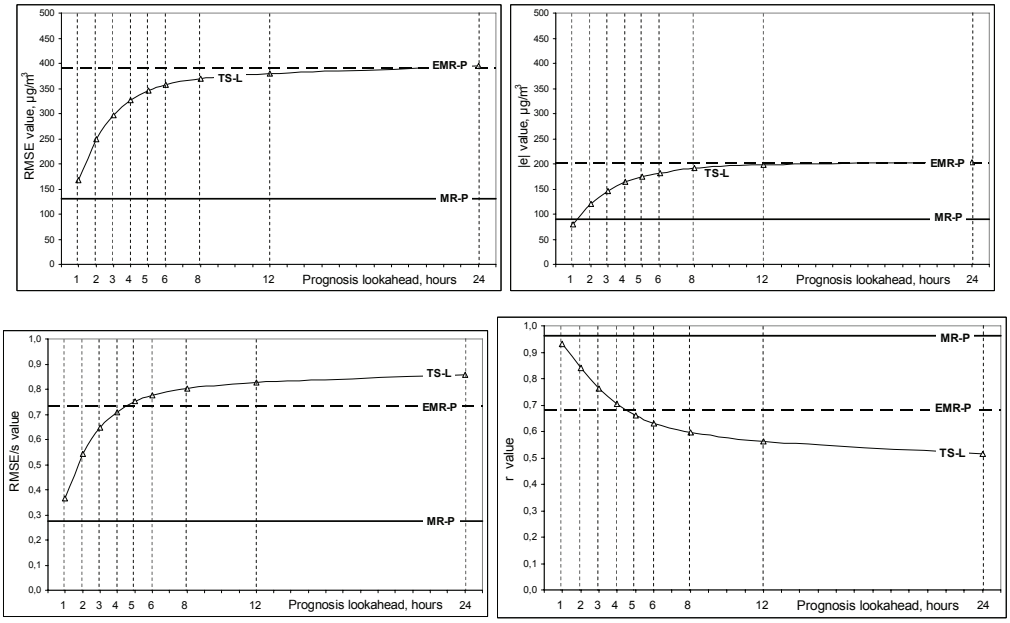

Fig. 5. Modelling errors of CO concentrations in dependence on the forecast horizon (Models TS-L, MR-P, EMR-P, Radom 2004–2008)

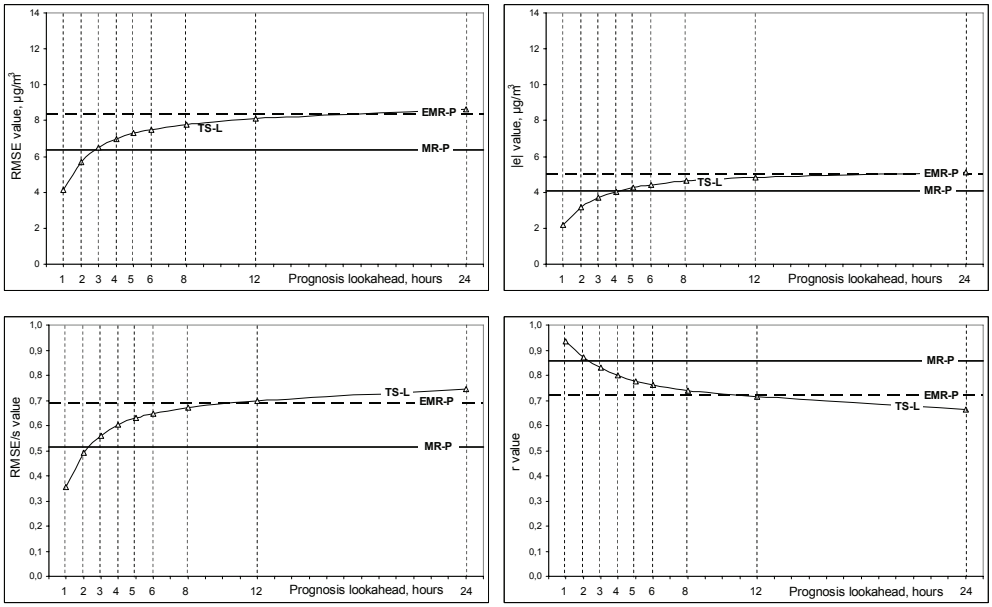

Fig. 6. Modelling errors of SO_2 concentrations in dependence on the forecast horizon (Models TS-L, MR-P, EMR-P, Radom 2004–2008)

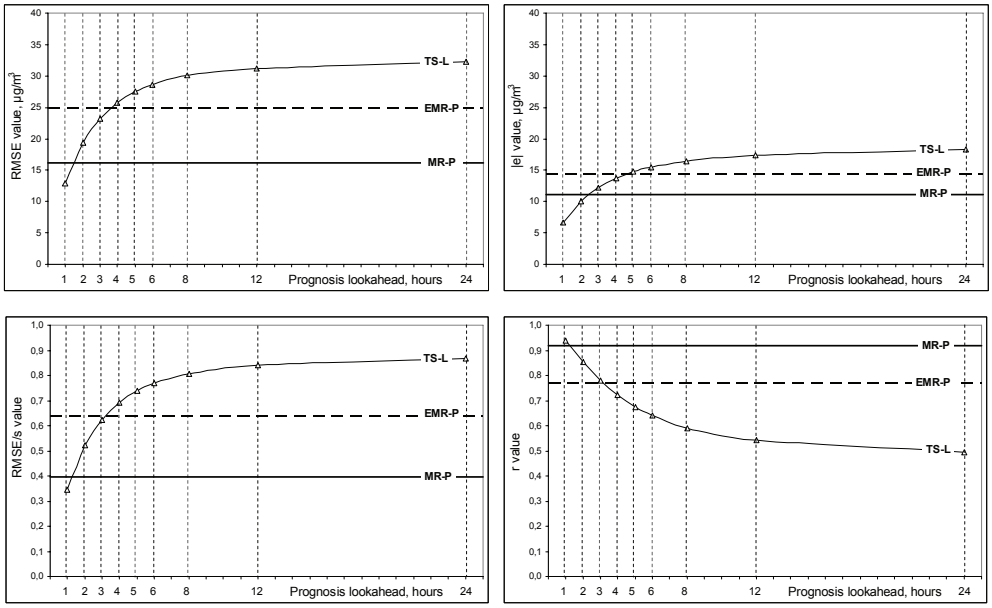

Fig. 7. Modelling errors of PM_{10} concentrations in dependence on the forecast horizon (Models TS-L, MR-P, EMR-P, Radom 2004–2008)

For ozone, TS-L models provide the highest prediction quality only in the first case of measuring gaps (Table 1). That recommendation does not depend on a considered type of error. The next cases in gaps should be modelled by regression models MR-P.

Table 1

The most accurate models of O_3 concentration for measuring gaps of various lengths

Error	Measuring gap length [h]	
	1	>1
RMSE	TS-L	MR-P
$\|e\|$	TS-L	MR-P
RMSE/s	TS-L	MR-P
r	TS-L	MR-P

For NO, TS-L models are not very accurate, even for the shortest forecast horizons. Instead, regression models MR-P are the most accurate, even compared to similar models for other pollutants. The correlation coefficient for this model is 0.954. Considering all types of measurement error, MR-P models should be recommended to predict NO concentrations from the first to the last case in a gap (Table 2).

Table 2

The most accurate models of NO concentration for measuring gaps of various lengths

Error	Measuring gap length [h]	
	1	>1
RMSE	MR-P	MR-P
$\|e\|$	MR-P	MR-P
RMSE/s	MR-P	MR-P
r	MR-P	MR-P

Table 3

The most accurate models of NO_2 concentration for measuring gaps of various lengths

Error	Measuring gap length [h]	
	1	>1
RMSE	TS-L	MR-P
$\|e\|$	TS-L	MR-P
RMSE/s	TS-L	MR-P
r	TS-L	MR-P

For NO_2 concentrations, TS-L models provide the highest prediction quality, but only in the first step of the forecast (Table 3). The next cases in gaps should be modelled by regression models MR-P.

For CO concentrations, for forecast horizons longer than 1 hour the most accurate type of model is MR-P (Table 4). For the first step, models TS-L can provide a higher prediction accuracy, but only when the mean absolute error $|e|$ is a criterion for prediction quality.

Table 4

The most accurate models of CO concentration for measuring gaps of various lengths

Error	Measuring gap length [h]	
	1	>1
RMSE	MR-P	MR-P
$\|e\|$	TS-L	MR-P
RMSE/s	MR-P	MR-P
r	MR-P	MR-P

Models TS-L give the best results in SO_2 concentrations modelling for the two first cases of a gap (Table 5). While considering mean absolute error $|e|$, this method proves to be the most accurate for measuring gaps up to 4 h. For longer gaps, MR-P models are most accurate.

Table 5

The most accurate models of SO_2 concentration for measuring gaps of various lengths

Error	Measuring gap length [h]				
	1	2	3	4	>4
RMSE	TS-L	TS-L	MR-P	MR-P	MR-P
$\|e\|$	TS-L	TS-L	TS-L	TS-L	MR-P
RMSE/s	TS-L	TS-L	MR-P	MR-P	MR-P
r	TS-L	TS-L	MR-P	MR-P	MR-P

For PM_{10}, the best results in the first step of prediction method always give models TS-L (Table 6). For the second step of the forecast, the error criterion decides on the recommendation. Analyzing the mean absolute error $|e|$, models of time series TS-L can be recommended. When comparing the values of other measurement errors, regression models of the type MR-P prove the best precision in the second step of the

forecast. These models provide the best prediction accuracy in every next step of the forecast.

Table 6

The most accurate models of PM_{10} concentration for measuring gaps of various lengths

Error	Measuring gap length [h]		
	1	2	>2
RMSE	TS-L	MR-P	MR-P
$\|e\|$	TS-L	TS-L	MR-P
RMSE/s	TS-L	MR-P	MR-P
r	TS-L	MR-P	MR-P

All the results indicate that models TS-L and MR-P provide the most accurate prediction for the data recorded at the air monitoring station in Radom. The accuracy of multivariate regression models exploring data from nearby monitoring station (EMR-P) is less than the accuracy of internal regression models (MR-P), for all pollutants. Therefore, this method cannot be recommended for prediction of measurement gaps, if other methods could be applied. Inaccuracy of EMR-P models can result from either particularly irregular primary emissions from local sources in the vicinity of the air monitoring station in Radom or the location of this station at the edge of the area monitored by the considered group of measuring stations.

NO is an exception among the pollutants for which models of the type MR-P provide the largest accuracy of approximation in the entire measurement gap. The time series method may be recommended for prediction in the first case of measuring gaps for concentrations of O_3, NO_2, CO. Exceptionally, the use of models TS-L should be considered to complete a larger number of cases: 1–2 first cases of gaps in time series of PM_{10} concentration, and 2–4 first cases of gaps for the time series of SO_2 concentrations. The choice of modelling method for subsequent cases in measurement gaps is obvious when only one measure of error is accepted as a criterion for selection of the modelling method. Accuracy evaluations resulting from the analysis of various measures of error are not divergent enough to reject any of these measures as an incorrect criterion.

5. CONCLUSIONS

- Specific methods of prediction should be recommended separately for each of the air pollutants, because there are big differences in possibilities of modelling of individual air pollution concentrations.

- In order to ensure optimum accuracy of prediction, the change of the modelling method with the lookahead increasing should be taken into account.
- Exceptionally for NO, the use a multivariate regression models MR-P is reasonable in the whole measuring gap.
- Concentrations of pollutants such as O_3, NO_2, CO, SO_2, PM_{10} can be effectively modelled using time series models but only for very short forecast horizons (1–4 h), then regression modelling methods appear to be more accurate.
- It should be only one measure of error assumed as a criterion for selection of the modelling method when missing data are completed. Then the recommendation of methods for the subsequent cases in a gap is simple.
- The above conclusions relate only to the data recorded at the air monitoring station in Radom, and only to the modelling methods applied as autonomous neural models. The analysis does not cover a comparison with other possible modelling techniques, including interpolation, which may be competitive for short gaps in time series.

ACKNOWLEDGEMENTS

This study was carried out within a research project of Czestochowa University of Technology No. BS-402-301/07/R. Some results were obtained earlier, in the project No. 1 T09D 037 30, funded by the research budget of the Polish Government.

REFERENCES

[1] PLAIA A., BONDI A.L., Atmos. Environ., 2006, 40 (38), 7316.
[2] GENTILI S., MAGNATERRA L., PASSERINI G., *An introduction to the statistical filling of environmental data time series*, [In:] *Handling Missing Data: Applications to Environmental Analysis*, G. Latini, G. Passerini (Eds.), Wit Press, Southampton, 2006, 1–27.
[3] Regulation of Ministry of Environment of 17 December 2008, *On the evaluation of levels of substances in the air*, Journal of Laws of the Republic of Poland, 2009, No 5, item 31.
[4] HOFFMAN S., *Treating missing data at air monitoring stations*, [In:] *Environmental Engineering*, L. Pawłowski, M.R. Dudzińska, A. Pawłowski (Eds.),Taylor and Francis Group, London, 2007, 349–353.
[5] HOFFMAN S., Environ. Eng. Sci., 2006, 23 (4), 603.
[6] HOFFMAN S., *Application of Neural Networks to Regression Modeling of Air Pollutants Concentrations*, Wydawnictwa Politechniki Częstochowskiej, Częstochowa, 2004 (in Polish).
[7] HOFFMAN S., JASIŃSKI R., *Missing Data Imputation into the Data Sets of Air Quality Monitoring Systems*, Wydawnictwo Politechniki Częstochowskiej, Częstochowa, 2009 (in Polish).