ARCHIVES OF ENVIRONMENTAL PROTECTION

vol. 34 no.3 pp. 133 – 142 2008

PL ISSN 0324-8461

COMPLETING MISSING DATA IN AIR MONITORING STATIONS USING DIURNAL COURSES OF REGIONAL POLLUTION CONCENTRATIONS

SZYMON HOFFMAN, RAFAŁ JASIŃSKI

Technical University of Częstochowa
Department of Chemistry, Water and Wastewater Technology
69 Dąbrowskiego St., 42-200 Częstochowa
szymon@is.pcz.czest.pl.; raphael@is.pcz.czest.pl

Abstract: Data sets gathered continuously in air monitoring systems are never entirely complete. The problem of missing data in monitoring measure series often has to be solved by modeling. A new method of air monitoring data modelling was tested in the paper. Regional diurnal concentration courses (RDCCs) were used as the main source of knowledge of predicted time series during specified days. The paper presents a comparison of predicted and measured diurnal concentration patterns of two frequently used parameters in air monitoring (PM_{10} and NO_2). The analysis was based on hourly time series of these air pollutants collected in a 3-year period at nine monitoring stations in the Lodz Region. It was shown that well determined regional diurnal concentration patterns could be useful to missing data modelling at the specified monitoring site. Improvement of modelling accuracy is possible after modification of modelling results by adding local difference vectors (LDVs), describing the specificity of the monitoring station.

Keywords: air monitoring data, missing data modelling, NO_2 concentration, PM_{10} concentration, diurnal courses

INTRODUCTION

The data sets gathered continuously in air monitoring systems are never complete (Hauck et al., 1999). It is possible to solve the problem of missing data in monitoring measurement series in many ways, using mainly two types of methods (Hoffman, 2007): 1) methods employing regression analysis, and 2) methods based on time-series analysis. Both groups of methods exploit classical statistics or other computational techniques based on artificial intelligence tools (Gardner and Dorling, 1998; Hadjiiski et al., 1999; Kolehmainen et al. 2001). Recently, modeling methods have been developed for requirements of local environmental assessments and authorities, as the tools of prediction (Karppinen et al., 2000; Hoffman, 2003; Plaia and Bondı, 2006; Hoffman, 2006). New modelling concepts, dedicated to specific applications, are still being requested and examined.

Every air monitoring site is situated in specific surroundings, which affects the measured pollution concentration level. Measuring point peculiarity may be mathematically projected by defining patterns of diurnal variation of pollution courses. It is suggested in the paper that well determined patterns should be useful to missing data modelling at the specified monitoring site.

An assessment of quality of missing data modelling with patterns of pollution concentration diurnal courses was the main purpose of our research. Modelling error was assumed as the criterion of modelling quality. The analysed data set was built of hourly averages, gathered in the long-term period at the air monitoring stations in the Lodz Region (Central Poland).

MATERIALS AND METHODS

Measuring data description

Long-term sets of hourly air monitoring data, gathered through the 3-year period (2004-2006) in Lodz Region (Central Poland) were used in the analysis. The data were received from the Voivodship Inspectorate for Environmental Protection in Lodz. These sets contained verified measurements from nine different monitoring stations: Gajew, Lodz-Widzew, Lodz-Center, Lodz-Zachodnia St., Pabianice-Polfa, Parzniewice, Piotrkow-Belzacka St., Radomsko-Sokola St., Zgierz-Center. A general description of the air monitoring sites was given in the tab. 1. The stations are representative of the different locations in the whole region.

Table 1.A general description of the air monitoring stations in Lodz Region.

Station	**Geographic coordinates**	**Area/station type**	**PM_{10} measure method**	**NO_2 measurement method**
Gajew	Longitude: 19°14'00" Latitude: 52°08'36"	rural/background	tapered element oscillation	chemiluminescence
Lodz-Center	Longitude: 19°27'19" Latitude: 51°46'04"	urban/background	tapered element oscillation	chemiluminescence
Lodz-Zachodnia St.	Longitude: 19°27'07" Latitude: 51°46'39"	urban/traffic	tapered element oscillation	chemiluminescence
Pabianice-Polfa	Longitude: 19°20'08" Latitude: 51 40'05"	urban/industrial	tapered element oscillation	chemiluminescence
Parzniewice	Longitude: 19°29'28" Latitude: 51°18'18"	rural/background	-	chemiluminescence
Piotrkow-Belzacka St.	Longitude: 19°40'23" Latitude: 51°24'25"	urban/background	tapered element oscillation	chemiluminescence
Radomsko-Sokola St.	Longitude: 19°26'31" Latitude: 51°03'50"	urban/background	gravimetry	chemiluminescence
Zgierz-Center	Longitude: 19°25'15" Latitude: 51°51'26"	urban/background	gravimetry	chemiluminescence
Lodz-Zachodnia St.	Longitude: 19°27'07" Latitude: 51°46'39"	urban/background	tapered element oscillation	chemiluminescence

Concentration courses of NO_2 and PM_{10} were chosen for the modelling study. Some descriptive statistics of the sets of NO_2 and PM_{10} concentrations are presented in the tab. 2. The table contains mean, minimum, maximum and standard deviation of concentration. The data are about 90% complete for both pollutants.

Table 2.Descriptive statistics of NO_2 and PM_{10} concentrations; Lodz Region, 9 stations, years 2004-2006.

Statistical parameter	Measure units	NO_2	PM_{10}
mean	$\mu g/m^3$	20.0	28.2
minimum	$\mu g/m^3$	0.0	0.0
maximum	$\mu g/m^3$	184.7	627.0
Standard deviation	$\mu g/m^3$	15.9	24.8
data completeness	%	89.2	90.1

Clusters of regional diurnal concentration courses

Separately for each day, regional diurnal courses of NO_2 and PM_{10} concentrations were designated by averaging the diurnal courses of concentrations observed for all air monitoring stations in the region on the respective day. The average courses obtained were treated in further analysis as regional diurnal concentration courses (RDCC). All RDCCs during the 3-year period were analyzed as separate cases by means of k-mean clustering. Five clusters for each pollutant were created. Cluster centres (means) represented 5 maximum distinct types of RDCC, observed for considered pollutants in Lodz Region.

Local difference vectors (LDV)

The next step in the analysis was a determination of 24-dimensional vectors of local difference (LDV). LDVs show how diurnal concentration courses of each local station differ from the regional pattern. It was assumed that LDV would enable improvement of modelling accuracy, because they describe specificities of particular stations.

Modelling of concentrations in the diurnal cycle

The considered modelling method is dedicated to the events (days), for which missing data appear in a single monitoring station, whereas data from the other stations are mostly complete with respect to a chosen pollutant. For each event separately, RDCC was calculated basing on diurnal concentration courses available from the other stations belonging to the regional monitoring network. Specified RDCC was corrected by adding LDV, determined for the station with missing values. LDVs are distinct for particular clusters, therefore at first the obtained RDCC was classified into one of five clusters, and then the proper LDV was added.

Modelling validation

The modelling error was estimated by comparison of modelled and real concentrations, separately for both types of models:

1. RDCC models, i.e. models using only regional diurnal concentration courses;
2. RDCC+LDV models, i.e. RDCC models modified by local difference vectors

At the beginning, the days with complete data from all nine monitoring stations were chosen in the whole 3-year period. Successively for each of these days, simulation of the lack of data was provided for the chosen monitoring station. Then, the artificially created missing data were regenerated with the use of both RDCC and RDCC+LDV models. This scheme was repeated for other monitoring stations.

For each day and each station genuine and regenerated daily concentration courses were compared, and two types of errors were calculated:

1. RMSE - root mean square error
2. lel - mean absolute error

The values of errors were averaged for the 3-year period, separately for NO_2 and PM_{10} concentration. The obtained averages of errors were used to estimate model accuracy.

RESULTS AND DISCUSSION

Results of cluster analysis of regional diurnal concentration courses

Five different clusters of regional diurnal concentration courses were obtained for both pollutants. The results of cluster analysis are shown in fig. 1 and fig. 2, for NO_2 and PM_{10} respectively. Curves plotted in the figures denote centres of specified clusters. It was assumed that each cluster represents the observations for the different levels of pollution emission/inflow in connection with specific weather conditions.

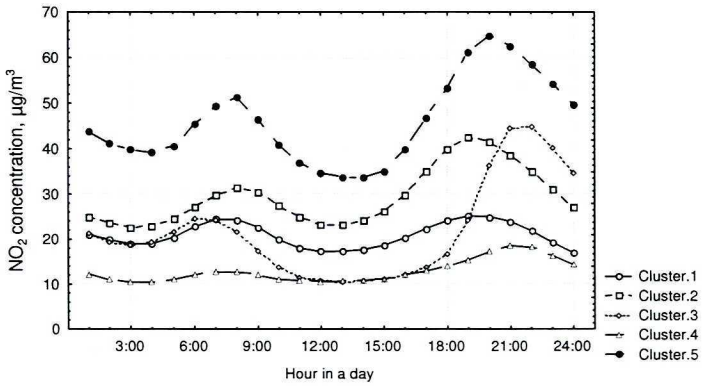


Fig. 1. Clusters of diurnal courses of NO_2 concentration (all monitoring stations in Lodz Region, 2004-2006).

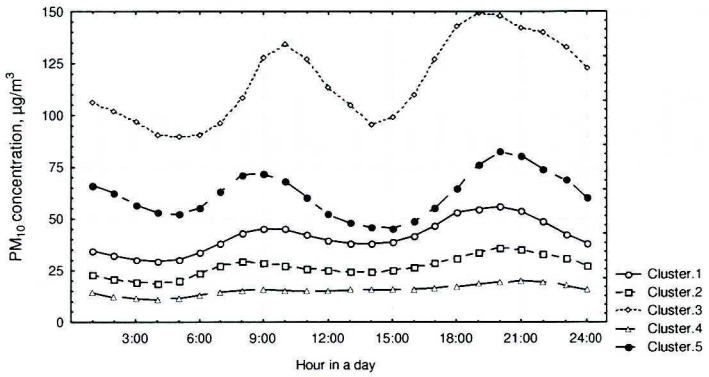


Fig. 2.Clusters of diurnal courses of PM_{10} concentration (all monitoring stations in Lodz Region, 2004-2006)

Some statistics of individual clusters are presented in tab. 3. Results show that the specified clusters represent very variable numbers of cases (days). In addition, their frequencies of occurrence also vary.

Table 3.Number of cases attributed to specified NO_2 and PM_{10} clusters (Lodz Region, 9 stations, 2004-2006)

Pollutant	Cluster number	Number of cases in the cluster	Percentage of cases in the cluster [%]
NO_2	1	204	19%
	2	386	35%
	3	13	1%
	4	432	39%
	5	53	5%
	Not classified	8	1%
PM_{10}	1	307	28%
	2	163	15%
	3	146	13%
	4	397	36%
	5	33	3%
	Not classified	50	5%

Results of local difference vectors calculation

Local difference vectors (LDVs) for every cluster of NO_2 and PM_{10}, denoted for different stations in the Lodz Region, are presented separately for both pollutants in the figs. 3 and 4. LDVs show the specificity of diurnal concentration variations at each local monitoring station in comparison with the regional pattern of concentration changes.

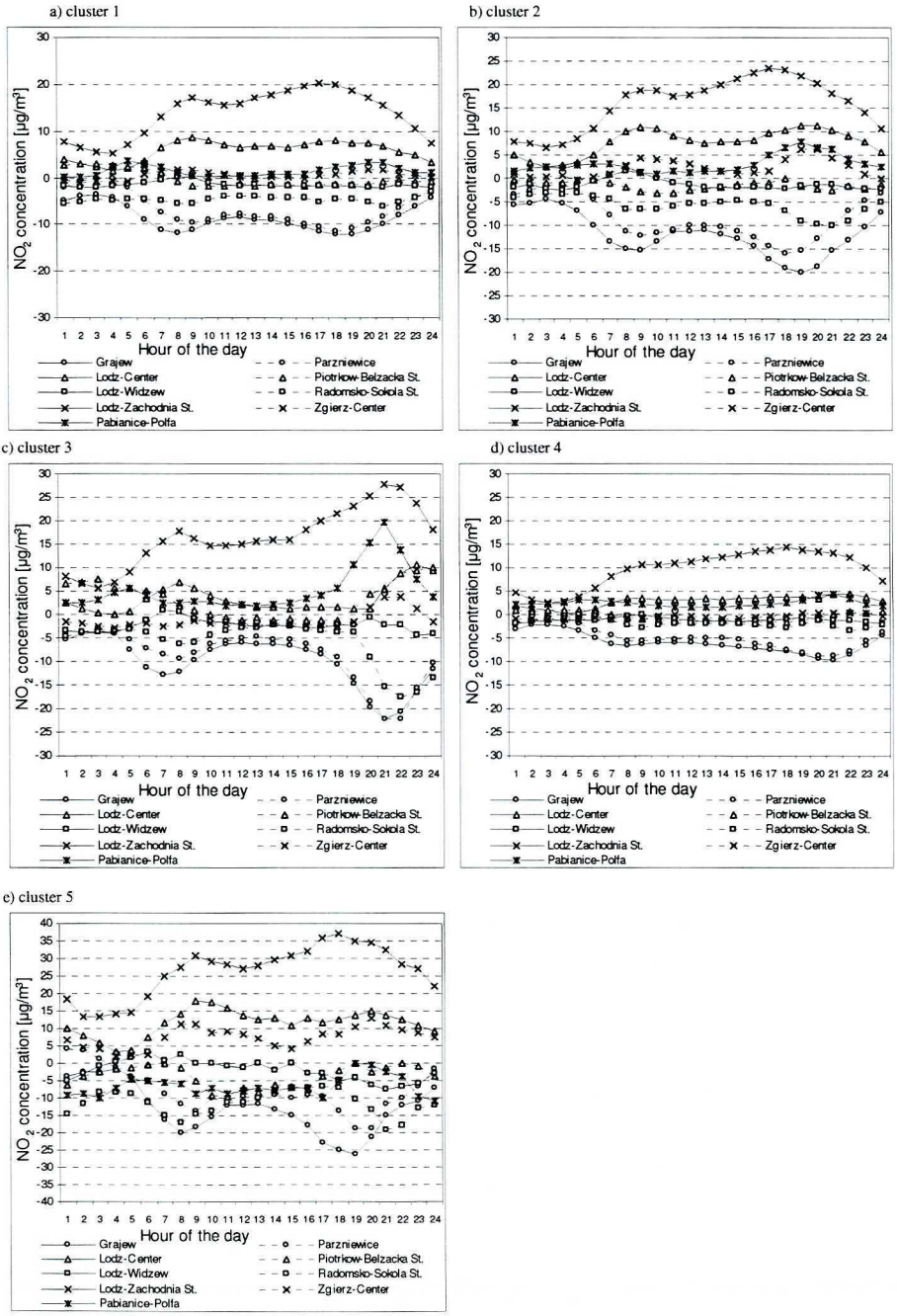
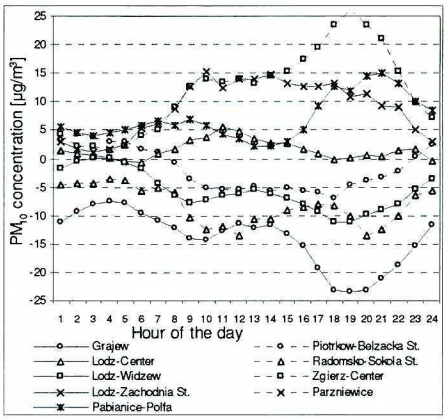
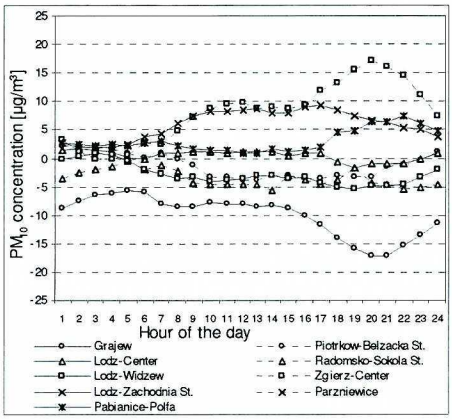


Fig. 3. Local difference vectors of NO_2 clusters, denoted for different stations in Lodz Region

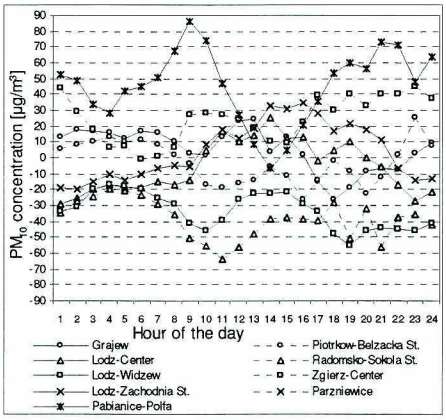
a) cluster 1



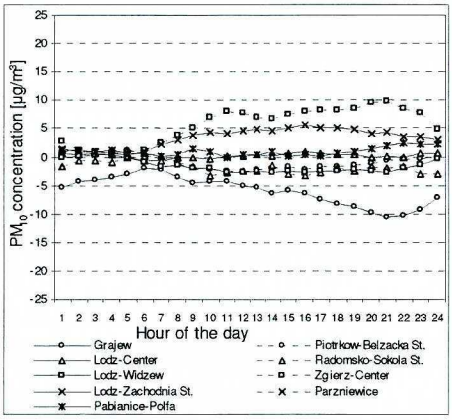
b) cluster 2



c) cluster 3



d) cluster 4



e) cluster 5

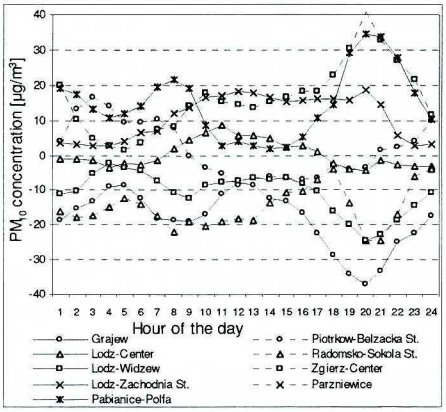


Fig. 4. Local difference vectors of PM_{10} clusters, denoted for different stations in Lodz Region

Results of modelling

The modeling quality was estimated for the days with entire diurnal concentration courses. The contribution of modeled courses in the whole 3-year period was distinct for individual stations, generally in the ranges: 47.4-88.8% for NO_2 and 43.3-84.9% for PM_{10}.

Table 4.Contribution of entire diurnal courses of NO_2 and PM_{10} concentration in the 3-year period for different air monitoring stations (Lodz Region, 2004-2006, 1096 days)

Monitoring station	NO_2 concentration		PM_{10} concentration	
	days	[%]	days	[%]
Gajew	879	80.2	635	57.9
Lodz-Widzew	971	88.6	866	79.0
Lodz-Center	973	88.8	890	81.2
Lodz-Zachodnia St.	969	88.4	930	84.9
Pabianice-Polfa	888	81.0	848	77.4
Parzniewice	912	83.2	-	-
Piotrkow-Belzacka St.	772	70.4	739	67.4
Radomsko-Sokola St.	520	47.4	475	43.3
Zgierz-Center	887	80.9	884	80.7

The values of modeling errors averaged for the 3-year period, separately for PM_{10} and NO_2 concentrations are shown in tabs. 3 and 4. The accuracies of RDCC models and adequate RDCC+LDV models are compared. The results confirm the assumption that LDV correction provides improved modelling accuracy. At all monitoring stations for both pollutants, errors of RDCC+LDV models are smaller than the errors of RDCC models. The biggest improvement of modelling quality is observed for RDCC+LDV models of NO_2 concentration courses (tab. 6). In some cases, over 50% decreases of RMSE or lel value were observed after LDV correction. The best results of PM_{10} modelling were obtained for the station Lodz-Center. NO_2 modelling was the most effective at the station Gajew (RDCC+LDV models) and at the stations Lodz-Widzew and Zgierz-Center (RDCC models).

Table 5.The average values of modelling errors of PM_{10} concentration courses (Lodz Region, 9 stations, 2004-2006)

Monitoring station	RDCC models		RDCC+LDV models	
	RMSE $\mu g/m^3$	lel $\mu g/m^3$	RMSE $\mu g/m^3$	lel $\mu g/m^3$
Gajew	17.7	14.6	12.6	10.0
Lodz-Widzew	10.2	7.8	8.6	6.4
Lodz-Center	8.0	5.8	7.9	5.8
Lodz-Zachodnia St.	11.3	8.6	9.0	6.6
Pabianice-Polfa	13.2	9.2	12.5	9.0
Parzniewice	-	-	-	-
Piotrkow-Belzacka St.	11.8	9.0	11.3	8.6
Radomsko-Sokola St.	14.5	11.4	12.4	9.4
Zgierz-Center	17.6	12.7	15.2	11.3

Table 6.The average values of modelling errors of NO_2 concentration courses (Lodz Region, 9 stations, 2004-2006)

Monitoring station	RDCC models		RDCC+LDV models	
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE $\mu g/m^3$	|e| $\mu g/m^3$
Gajew	12.4	10.7	6.6	5.2
Lodz-Widzew	7.4	6.1	7.2	5.7
Lodz-Center	9.0	7.4	7.5	6.1
Lodz-Zachodnia St.	17.1	15.0	9.0	7.3
Pabianice-Polfa	10.3	8.3	10.1	8.2
Parzniewice	12.0	10.1	8.4	6.5
Piotrkow-Belzacka St.	8.6	6.8	8.4	6.6
Radomsko-Sokola St.	9.7	7.9	8.1	6.4
Zgierz-Center	7.4	6.0	7.2	5.8

Table 7. The maximum values of modelling errors of NO_2 concentration courses (Lodz Region, 9 stations, 2004-2006)

Monitoring station	RDCC models		RDCC+LDV models	
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE $\mu g/m^3$	|e| $\mu g/m^3$
Gajew	31.4	29.4	30.3	27.8
Lodz-Widzew	29.4	24.1	31.4	24.1
Lodz-Center	41.3	37.6	32.4	26.5
Lodz-Zachodnia St.	52.5	45.2	30.0	27.0
Pabianice-Polfa	78.8	76.3	72.9	69.7
Parzniewice	37.2	32.2	31.6	25.6
Piotrkow-Belzacka St.	34.1	25.1	29.8	22.4
Radomsko-Sokola St.	35.0	34.1	34.2	23.1
Zgierz-Center	27.4	25.5	22.8	20.6

Table 8. The maximum values of modelling errors of PM_{10} concentration courses (Lodz Region, 9 stations, 2004-2006)

Monitoring station	RDCC models		RDCC+LDV models	
	RMSE $\mu g/m^3$	|e| $\mu g/m^3$	RMSE $\mu g/m^3$	|e| $\mu g/m^3$
Gajew	86.2	103.6	83.3	101.6
Lodz-Widzew	97.1	66.6	87.9	59.8
Lodz-Center	66.8	57.6	55.3	47.4
Lodz-Zachodnia St.	89.7	85.0	42.2	39.2
Pabianice-Polfa	195.6	150.8	158.9	114.4
Parzniewice	-	-	-	-
Piotrkow-Belzacka St.	125.4	124.4	89.6	85.7
Radomsko-Sokola St.	95.8	101.2	77.8	61.0
Zgierz-Center	106.6	85.1	89.8	66.6

The maximum values in each category of modelling errors were presented in tab. 7-8. The maximum errors show the modelling quality in the most extreme situations (episodes). The maximum errors are distinctly higher than the average ones. This fact forces us to conclude that episode modelling by new method does not give satisfactory results. Especially high errors were observed at the station Pabianice-Polfa, which is an industrial area.

CONCLUSIONS

A new method of air monitoring data modelling was tested in this paper. The temporal diurnal courses of regional concentrations were used as the main source of knowledge of predicted time series on selected days. The paper presents a comparison of predicted and measured diurnal concentration patterns of PM_{10} as well as NO_2. Both the measured and predicted data includes time series of these air pollutants for the 3-year period in the Lodz Region. It is suggested that well determined regional diurnal concentration patterns should be useful to missing data modelling at the specified monitoring site.

The results allow the following general conclusions:

1. Regional patterns of the diurnal courses of pollutant concentrations may be applied to missing data modelling in air monitoring systems.
2. Improvement of modelling accuracy is possible after modification of modelling results by adding local difference vectors, describing the specificity of the monitoring station.
3. Modelling of the highest concentration episodes by new method does not give satisfactory results.

The above results were obtained for NO_2 and PM_{10} modelling. It is expected that modelling of other pollutants levels will confirm the statements expressed above.

Acknowledgements

This work was carried out within the framework of the research project number 1 T09D 037 30, funded by research budget of Polish Government for the years 2006-2008.

REFERENCES

[1] Gardner M.W., Dorling S.R., 1998. Artificial neural networks (the multilayer perceptron) – a review of applications in the atmospheric sciences. Atmospheric Environment 32, 2627-2636.

[2] Hadjiiski L., Geladi P., Hopke P., 1999. A comparison of modeling nonlinear systems with artificial neural networks and partial least squares. Chemometrics and Intelligent Laboratory Systems 49, 91-103.

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

[4] Hoffman S., 2003. Regression modelling of ground level ozone concentration. In L. Pawłowski, M.R. Dudzińska, A. Pawłowski, (Eds.), Environmental Engineering Studies. Polish Research on the way to EU. Kluwer Academic/Plenum Publishers, New York, 53-60.

[5] Hoffman S., 2006. Short-Time Forecasting of Atmospheric NO_x Concentration by neural networks. Environmental Engineering Science 23(4), 603-609.

[6] Hoffman S., 2007. Treating missing data at air monitoring stations. In L. Pawłowski, M.R. Dudzińska, A. Pawłowski (eds.),„Environmental engineering", Taylor & Francis Group, London, 349-353.

[7] Karppinen A., Kukkonen J., Elolähde T., Konttinen M., Koskentalo T., 2000. A modelling system for predicting urban air pollution: comparison of model predictions with the data of an urban measurement network in Helsinki. Atmospheric Environment 34, 3735-3743.

[8] Kolehmainen M., Martikainen H., Ruuskanen J., 2001. Neural networks and periodic components used in air quality forecasting. Atmospheric Environment 35, 815-825.

[9] Plaia A., Bondı A.L., 2006. Single imputation method of missing values in environmental pollution data sets. Atmospheric Environment 40, 7316–7330.

Received: September, 2007; accepted: June, 2008.