Optimizing the prediction models of the air quality state in cities

J. Skrzypski, E. Jach-Szakiel & W. Kamiński
Faculty of Process and Environmental Engineering,
Technical University of Lodz, Poland

Abstract

The aim of research was to optimize the neural networks models for predicting the classes of air quality state. This model was constructed and tested on the basis of the data gathered in Lodz, a city localized in the middle of Poland. Models were tested in relation to mean daily dust concentration (PM_{10}) as well as maximal daily values. In each case, 5 air quality classes were distinguished. Air quality in each day was classified with respect to the meteorological conditions. Two models were built: two artificial neural networks (ANNs) were used, namely the MLP and the RBF. Optimization relied on determining the value of the following optimal value of the following ratio of the training set to the testing set, neuron number, number of input vectors after the PCA dimension reduction procedure.

Results of the modelling are satisfactory. The error for predicting the air quality classes was generally smaller than 13%. In relation to predicting air quality classes, mean daily PM_{10} concentrations, better results were obtained with the RBF model containing 5 neurons. The RBF model for maximal daily PM_{10} concentrations generates a classification error of about 10,7%, and MLP model generates an error of 14,9%.

Keywords: artificial neural network, classification, PM_{10},

1 Introduction

A feature of the climate in large cities is the unsatisfactory condition of the atmospheric air due to human activity. Unneeded interactions are due to the functioning of the developed road transport, many branches of industry located in a small area as well as realization of the life and economic needs such as

compartments heating [1]. As an effect the danger of high concentration of air pollution may arise, which may have negative influence on the health and even the life of many citizens. In managing the air quality for the large cities the diagnostic and prognostic models play a vital role. The development of such tools is a complicated task. The process of forming the deposition fields are dependent on many various factors with strong variability of parameters. Moreover, difficulties result from strong dynamics of atmosphere as well as nonlinearities and incomplete recognition of atmospheric processes and imperfections in the monitoring network [2]. Taking all elements and conditions into account is impossible. In practice, the range of available data is limited, therefore farfetched simplifications are applied in models, especially numerical models.

Artificial Neural Networks (ANNs) are a proven tool for deposition concentration fields description. Due to their ability to generalize dependencies and reduce impact of secondary factors ANNs allow one to use incomplete measurement data burden with errors. ANN are especially useful in forecasting pollution level in cities, where complexity of problem preclude different methods of modeling. Many works with similar approach can be stated [3–7].

The atmosphere monitoring system in Poland was greatly improved. The higher quality and extended range of atmosphere monitoring allows a more precise determination of air quality characteristics and effect-reason analysis. Possibility of determination and analysis the daily concentration charts is very valuable. In previous years this was not possible due to the generation of less reliable results and the only results available were 24-hour mean air pollution concentrations. Recognition of the scale, specific and conditions of the 24-hours variability in the concentrations significantly improves basis of the air protection management, especially predicting the mean and maximal concentrations.

2   Territory and scope of research

The testing ground for our research was the city of Lodz (population: 800 000) located in the center of Poland. In the present paper the results of prediction concern the dust pollution (PM_{10}) in the densely build city center, where state of air quality is determined with the pollution emission originated with low emitters (individual house furnaces). Results obtained in the period starting from 1 January 2004 and ending 28 February 2006 at an automatic monitoring station located in the city center were used.

A typical feature of the daily run of the dust concentration PM_{10} during the winter season is presence of two distinct maxima: lower from 9^{00}-10^{00} and higher from 17^{00}- 19^{00} (Fig. 1). Daily minimal dust concentration occurs from 04^{00}-06^{00}. At the station located amongst housing estates the amplitude of daily concentrations is only 18 μg/m^3, due to small values of daily maxima. At the stations located in the city center a decrease in the concentration (by the 20 μg/m^3 compared to the afternoon maxima) at midday can be clearly seen. The period of the relatively high PM_{10} concentrations lasts – in winter season – more than 12 hours (from 08^{00} to 21^{00}).
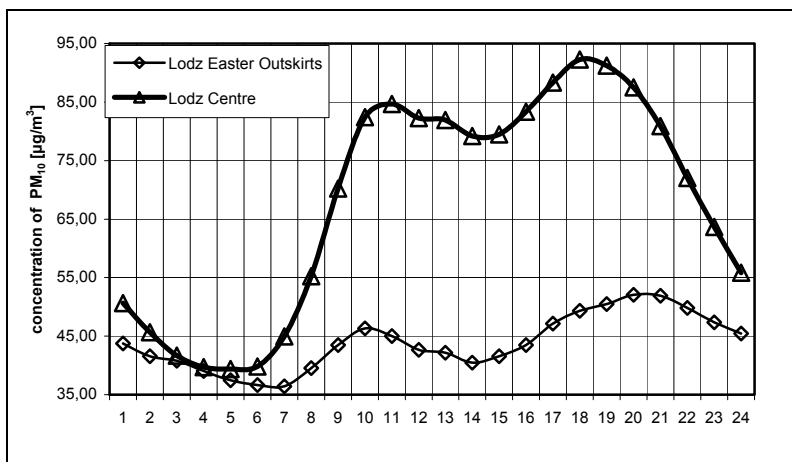
Figure 1:   Daily changes of mean values of PM_{10} dust concentration in winter seasons 2004-2006 (I, II, XII) for monitoring stations localized in Lodz.

PM_{10} concentration during the summer season is significantly lower than during the winter. Even in the city center the mean daily maximal values are twice lower than the winter maxima. Main summer maximum occurs only at 20^{00}-22^{00} hours, and secondary maximum at 06^{00}-08^{00} hours. As a consequence the near-noon period of decreased concentrations during summer lasts a few hours longer than during winter. In this period PM_{10} concentrations in cities can be as high as 30-35 $\mu g/m^3$. Daily minima on all city stations are as low as 25 $\mu g/m^3$. These minima are registered during night hours (02^{00}-04^{00}).

3   Model optimization procedure

To forecast air quality the determination of the precise concentration value is not always needed. To warn citizens on high air pollution concentration hazards of ranges (classes) of forecasted values of concentrations were determined. The aim of our research was to optimize the neural networks models for the predicting classes of air quality state.

Measured data of the dust pollution concentration were grouped into 5 class ranges depending on the concentration value. The class width was declared according the effective standards of the air quality (and multiplication factor of these standards, e.g. 50%, 100%, 150% of the standard values). The size of the created range of classes and number of cases for each class are shown in Tables 1 and 2. A similar approach was presented by Perez and Reyes [8].

Within the frames of present work two series of air quality classification were made. The first classification comprises distributions based on maximal daily value (Table 1) of the PM_{10} concentration between hours 08^{00}-20^{00}. The second classification was based on mean daily values (Table 2) of the $PM_{10.}$

concentration. Additionally, for each classification meteorological conditions for given measurement day were assigned.

To classify air quality in the aspect of state of dust pollution two types of artificial neural networks were used. First was artificial neural network of multi-layer perceptron (MLP) type with one layer of hidden neurons. In the second state of simulation neural networks of the Radial Basic Function (RBF) type was used. Additionally, a Principal Component Analyses (PCA) procedure was used to decrease dimension of input vectors (number of meteorological data).

PCA methods allow us to increase number of neurons in the hidden layer keeping constant total number of weights in neural networks. To have comparable results for each version the number of weights in neural networks should be similar.

Table 1:   Characteristics of input data used in air quality classification in aspect of maximal PM_{10} concentration during day.

Parameter	Classes of PM_{10} concentration[$\mu g/m^3$]				
	I	II	III	IV	V
Range of classes [$\mu g/m^3$]	1-49	50-99	100-149	150-199	>200
Number of cases	25	89	68	33	25
Mean value of maximal concentration during day	40,9	76	123,6	167,9	284,2
Mean daily sum of rainfall [mm]	1,19	0,75	0,47	0,16	0,32
Mean daily sum of radiation [W/m^2]	366,3	503,1	775,9	819,9	1257
Mean daily wind velocity [m/s]	4,6	3,6	3,1	2,7	2,36
Mean daily vertical velocity [m/s]	-0,01	-0,07	-0,13	-0,09	-0,12
Mean maximal daily temperature [˚C]	1,9	0,3	-0,6	-2,5	-4,07
Mean daily minimal temperature [˚C]	-1,3	-3,3	-4,3	-6,8	-10,8
Mean daily atmospheric pressure [hPa]	985	989	990	990	994

Table 2:   Characteristics of input data used in air quality classification in aspect of mean PM_{10} concentration during day.

Parameter	Classes of PM_{10} concentration [$\mu g/m^3$]				
	I	II	III	IV	V
Range of classes [$\mu g/m^3$]	1-25	25,01-50	50,01-75	75,01-100	>100
Number of cases	16	82	59	52	31
Mean value of maximal concentration during day	20,8	38,6	61,5	85,6	153
Mean daily sum of rainfall [mm]	1,36	0,86	0,45	0,38	0,13
Mean daily sum of radiation [W/m^2]	296,1	509,0	633,2	864,8	1173,1
Mean daily wind velocity [m/s]	4,7	3,7	3,1	2,8	2,2
Mean daily vertical velocity [m/s]	0,01	-0,07	-0,09	-0,14	-0,16
Mean maximal daily temperature [˚C]	1,8	1,0	-0,9	-1,8	-3,8
Mean daily minimal temperature [˚C]	-1,2	-2,4	-4,5	-6,4	-9,8
Mean daily atmospheric pressure [hPa]	984	989	990	990	992

Neuron model optimization relies on determination:

- ratio of the size of the training file against test file for the model (4 versions: test file was 17%, 20%, 25% and 33% of the all data file);
- number of neurons hidden in the MLP model (3 versions: 3, 5 or 7 neurons);
- number of input vectors characterizing meteorological conditions in the form of real meteorological conditions or orthogonal components obtained thanks to PCA;
- information on air quality in previous day.

Table 3:   Characteristics of input vectors in the MLP calculation of the air quality.

Calculation case	ANN type	Information on meteorological conditions	Information on state of air pollution	Number of hidden neurons	Number of weights
1.		7 meteorological vectors	Max value of daily and previous day concentration	5	51
1a.	MLP	5 orthogonal components		6	49
1b.		3 orthogonal components		8	49
2.		7 meteorological vectors	Max value of daily concentration; max and mean values for the previous day	5	56
2a.	MLP	5 orthogonal components		6	55
2b		3 orthogonal components		8	57
3.		7 meteorological vectors	Mean value of daily and previous day concentration	5	51
3a.	MLP	5 orthogonal components		6	49
3b.		3 orthogonal components		8	49
4.		7 meteorological vectors	Mean value of daily concentration; max and mean value for previous day	5	56
4a.	MLP	5 orthogonal components		6	55
4b		3 orthogonal components		8	57

On the first stage of optimization the size of test file and number of neurons/centers in the network was determined. Calculations shown that error of forecasting the air quality class is the smallest if:

- test file size is 20% of available data file size;
- in the network there are 5 hidden neurons in the MLP model and 5 centers in the RBF model.

On the second stage we concentrated on network input vector structure optimization in the aspect of meteorological data and data on previous day PM_{10} concentration (Tables 3 and 4).

Table 4:   Characteristics of input vectors in the calculation of the air quality classification with the ANN of the RBF type.

5.		7 meteorological vectors	Max value of daily and previous day concentration	5	50
5a.	RBF	5 orthogonal components		6	48
5b.		3 orthogonal components		8	48
6..		7 meteorological vectors	Max value of daily concentration; max and mean values for the previous day	5	55
6a.	RBF	5 orthogonal components		6	54
6b		3 orthogonal components		8	56
7.		7 meteorological vectors	Mean value of daily and previous day concentration	5	50
7a.	RBF	5 orthogonal components		6	48
7b.		3 orthogonal components		8	48
8.		7 meteorological vectors	Mean value of daily concentration; max and mean value for previous day	5	55
8a.	RBF	5 orthogonal components		6	54
8b		3 orthogonal components		8	56

4   The results of classification

All calculations were made for two types of neural networks (MLP and RBF). Air class forecasting in the aspect of the mean daily and maximum daily PM_{10} concentration has been done separately. In some cases model was tested with the PCA method with changes in the input vectors dimensions.

The results of air quality class forecasting were determined in a three-stage arrangement:

Stage I– good results: difference between calculated and real class is equal to 0 (full conformity of the forecast with reality),

Stage II– acceptable results (correct): difference between calculated and real class is +/-1 (forecast differs from reality by one concentration class),

Stage III– no acceptable results (erroneous): difference between calculated and real class is greater than +1 and smaller than –1 (forecast differs from reality by more than one concentration class).

All calculations have been made for the training and test series. In training series the rate of good results obtained with the MLP neural network was from 66,7% (for 4th case) to 49,5% (for 2b case), and for RBF neural network such rate was from 56,8% (for 8 case) to 46,9% (for 6a case). The rate of erroneous results in the training series was:

-    for the MLP network minimal rate of erroneous results was 0 (for calculation case 3) and maximal rate was 6,8% (for case 1a),

-    for the RBF network minimal rate of erroneous results was 1 (for calculation cases 7 and 8) and maximal rate was 7,8% (for case 6a).

Results obtained in test series were discussed in detail. Results for all cases are shown in Tables 5-8.

In the first series (calculation cases 1, 3, 5, 8) of calculation input vector consisted of information on the following:

- on air pollution state (1 vector) in the previous day; It was maximal PM_{10} concentration value during the day or mean daily PM_{10} concentration value, depending on the case,
- on meteorological condition in the given day (7 components of the input vector).

In this series the best results were obtained using RBF neural network (calculation case 7). In this case the highest rate of good results (61,7%) and 2% of erroneous results was obtained for the test series (Table 8).

Table 5:   Results of the forecasting the air quality class for the daily maximal PM_{10} concentration in the Lodz's city center during winter seasons 2004-2006 using ANN of MLP type.

Calculation case	Rate (%) of results in test series		
	good	acceptable	erroneous
1.	46,8	42,6	10,6
1a.	40,5	46,8	12,7
1b.	34,1	57,4	8,5
2.	38,3	55,3	6,4
2a.	38,3	44,7	17,0
2b.	51,1	38,3	10,6

In the second series (calculation cases 2, 4, 6, 8) the information on PM_{10} pollution state was extended in such a way that input vector contained maximal and mean daily concentration values for the previous day. The meteorological vector remained unchanged.

In this series the best results were obtained using ANN of RBF type (calculation case 8). For the test series the high rate of good results (57,4) and zero rate of erroneous results was obtained (Table 8).

Table 6:   Results of the forecasting the air quality class for the mean daily PM_{10} concentration in the Lodz's city center during winter seasons 2004-2006 using ANN of MLP type.

Calculation case	Rate (%) of results in test series		
	good	acceptable	erroneous
3.	59,6	38,3	2,1
3a.	44,7	46,8	8,5
3b.	31,9	57,4	10,7
4.	40,4	55,3	4,3
4a.	38,3	46,8	14,9
4b.	42,6	44,7	12,7

In the third series (calculation cases: 1a, 2a, 3a, 4a, 5a, 6a, 7a, 8a), the meteorological vector was modified. Vector, consisting of 7 values was

transformed to 5 orthogonal vectors. Results obtained in this series are not satisfactorily, because of the decrease in the rate of good results and significant increase in the rate of erroneous results.

Table 7:   Results of the forecasting the air quality class for the maximal daily PM_{10} concentration in the Lodz's city center during winter seasons 2004-2006 using ANN of RBF type.

Calculation case	Rate (%) of results in test series		
	good	acceptable	erroneous
5.	48,9	46,8	4,3
5a.	40,4	48,9	10,7
5b.	51,1	42,6	6,3
6.	53,2	40,4	6,4
6a.	44,7	46,8	8,5
6b.	40,4	48,9	10,7

In the fourth series of calculations (calculation cases 1b, 2b, 3b, 4b, 5b, 6b, 7b, 8b) meteorological vector was transformed to three orthogonal vectors. Significantly better results compared to earlier calculation series were obtained only for cases 2b, 4b and 5b (Table 5-7). In each case the rate of good results was higher than 50 %. The rate of erroneous results was from 6% to 12 %.

Table 8:   Results of the forecasting the air quality class for the mean daily PM_{10} concentration in the Lodz's city center during winter seasons 2004-2006 using ANN of RBF type.

Calculation case	Rate (%) of results in test series		
	good	acceptable	erroneous
7.	61,7	36,2	2,1
7a.	46,8	44,7	8,5
7b.	38,3	57,4	4,3
8.	57,4	42,6	0,0
8a.	44,7	51,1	4,2
8b.	44,7	46,8	8,5

5   Conclusions

The possibility of forecasting of the air quality classes in cities using neural network modeling was analyzed. Obtained results are satisfactory and thus confirming the usability of neural networks for solving such a class of problems. Taking into account all calculation cases (24) for both types of neural networks, mean error in forecasting the air quality class is 8%; greatest error is 17%, the smallest – 0%.

Simultaneously, the results of our work indicate the necessity of model optimization considering various conditions, especially in relation to number and type of input vectors. Comparing the optimization results of the air quality classification models for each calculation case, the following conclusions can be made.

- The greatest rate of good results was obtained in calculation case 7, when air quality class was forecasted in respect of mean daily PM_{10}, concentration, using RBF type network without implementing the PCA method considering mean daily PM_{10}, concentration for the previous day (Table 8),
- The lowest rate of erroneous results was obtained in calculation case 8, when air quality class was forecasted in respect of mean daily PM_{10} concentration using RBF network without implementing the PCA method considering mean daily PM_{10} concentration for previous day and maximal concentration value for previous day (Table 8),
- Using the same type of neural network better results were obtained for calculation cases using mean daily PM_{10} concentration.
- Only in three cases results were better with PCA method (calculation cases 2b, 4b i 5b; Tables 5-7),
- Using the same input vectors in the various neural networks slightly better results were obtained using RBF type network,
- Apart from network type and calculation case slightly better results of the air class forecasting were obtained in relation to the mean daily value of PM_{10} concentration; slightly worse results of air quality class forecasting for maximal daily PM_{10} concentration may result due to the small number of the highest concentration cases; it may be a barrier on the network training stage.

References

[1] Mayer H. Air pollution in cities, *Atmospheric Environment*, 33, pp. 4029-4037, 1999.
[2] Skrzypski J., Analysis and modelling of the system of air pollution concentration fields in big cities, In: *Environmental Engineering Studies*, ed. Pawłowski at al., Kluwer Academic/Plenum Publishers, New York, pp. 69-78, 2003.
[3] Lu W.Z., Wang W.J., Wang X.K., Yan S.H., Lam J.C., Potential assessment of neural network model with PCA/RBF approach for forecasting pollutants trends in Mong Kok urban air, Hong Kong, *Environmental Research* 97 pp. 79-87, 2004.
[4] Niska H., Hiltunen T., Karppinen A., Ruuskanen J., Kolehmainen M., Evolving the neural network model for forecasting air pollution time series, *Engineering Applications of Artificial Intelligence*, 17, pp. 159-167, 2004.
[5] Kukkonen M, Niska H., Dorling S., Chatterton T., Foxall R., Caweley G., Extensive evaluation of neural network models for prediction of NO_2 and PM_{10} concentrations, compared with deterministic modeling and

measurements in central Helsinki, *Atmospheric Enviroinment*, 37, pp. 4539-4550, 2003.

[6] Vitti P., Liuti G., Di Genova P., Atmospheric urban pollution: application of an artificial neural network (ANN) to the city of Perugia, *Ecological Modelling*, 148, pp. 27-46, 2002.

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

[8] Perez P., Reyes J., PM_{10} forecasting in Santiago, Chile, WIT Press, *Air Pollution XIII*, Ed. C.A. Brebbia, pp.33-38, 2005