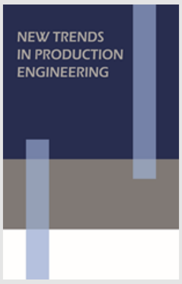
sciendo

DOI 10.2478/ntpe-2019-0055



2019
Volume 2
Issue 1
pp. 515-523

Application of neural networks to the prediction of gas pollution of air

Małgorzata Pawul
AGH University of Science and Technology, **Poland**

Date of submission to the Editor: 05/2019
Date of acceptance by the Editor: 08/2019

INTRODUCTION

Air pollution is currently one of the main environmental protection issues. Air pollution occurs not only in the cities, but also on rural areas. According to the WHO data, 91% of the global population live on the areas affected by poor air quality, and that has negative consequences for health conditions (www.who.int, 2019). A number of actions can be carried out to improve air quality, including making changes in the production methods or the replacement of heating systems for more ecological ones. Preventive measures are also introduced in the cities, with the intention to stop excessive air pollution, within several days ahead. An example of such activities consists in offering free public transportation, with the expectation of traffic reduction causing lowering of air pollution rates. It is necessary to design detailed air pollution forecasts to be able to plan such actions. Particulate matter is the type of pollution occurring most often in Poland and that factor often exceeds the allowed concentration rates. However, air pollution forecasts should also be drafted for gases.

We can distinguish deterministic and statistical models of air pollution among the methods of air pollution forecasts. The deterministic models allow for adopting forecasts for a whole area under consideration, also in the absence of measured values on the area (smog.imgw.pl, 2019). A forecast is designed on the basis of mathematical models of the spread of pollution. An example of such a forecasting system for air pollutant distribution is the FAPPS system, designed for the Małopolska Region (Hajto et al., 2012). Statistical methods, on the other hand, do not require any large computing capabilities as it is the case with deterministic models, and they are drafted on the basis of historical data. The models designed on the basis of machine learning, including artificial neural networks, adopted to forecast the levels of a specific air pollutant at the given place (point) and time, are the examples of statistical models. The disadvantage of those models is that they cannot be used for forecasting the situation on a wide area.

Contact address: e-mail: pawul@agh.edu.pl

A number of works have been dedicated to air quality modelling. Some articles refer to prediction of particulate matter and gas pollution (Abderrahim et al., 2016; Kukkonen et al., 2003; Pawul and Śliwka, 2016). A review of the research in that field was published by Rybarczyk and Zalakeviciute (Rybarczyk and Zalakeviciute, 2018). The authors indicated that the research methods concerning PM10 and PM2.5 pollution dominated among the pollution predictions, with the application of machine learning methods. A number of studies were also dedicated to forecasting nitric oxides and ozone and few dedicated to sulphur dioxide or carbon monoxide. When considering the frequency of applying forecasting methods, artificial neural networks are just behind the Ensemble Learning Methods.

In our research, we concentrated on the application of neural networks to the prediction of air pollution levels. However, it is worth mentioning here that neural networks are also considered in resolving other issues relating to environmental protection (Haupt et al., 2009; Kwiecień and Pawul, 2012; Tadeusiewicz and Dobrowolski, 2004).

METHODOLOGY OF RESEARCH

The principles of operation of artificial neural networks

Among many methods and algorithms of artificial intelligence (AI), artificial neural networks (ANNs) have the ability of representing a number of complex functions. Their structure and the principles of operation allow us to apply them to resolve the problems in which the function describing a given process is not known (i.e. the function of relation between input and output data). ANNs consist of units (neurons) that process specific data. Each neuron participating in the network is connected to other neurons by the links with specific parameters (or so-called synaptic weights), and those links are changing during the training process. Based on the current state of neuron activation and input signals, the signal to be sent by the neuron to the other neurons (or network nodes) is calculated. During such a transmission, the signal is either weakened or strengthened, depending on the connection characteristics (Tadeusiewicz, 1993).

In each neural network, the neurons are grouped to create layers. A layer is composed of a collection of neurons where each of them contains the same set of input signals and its own and separate weight vector. The neural networks are made of several layers: the input layer used for the introduction of input date to the network, the output layer determining the final solution, and the hidden layers that process signals in the way allowing to extract certain intermediate data that are necessary to determine the final solution. Within one layer, the neurons are connected with the neurons of the subsequent layer, in whole or in part. Besides, feedback connections to preceding layers may also exist. It is necessary to determine the number of layers and the numbers of neurons in each layer to define multilayer networks (Tadeusiewicz, 1993).

There are many types of networks that are different in respect of their structures and operating principles. Non-linear multilayer networks, or the so-called multilayer perceptrons, are quite common among neural network structures (Tadeusiewicz, 1993). The determination of the proper number of layers and of neurons in each layer is a very important stage of the whole network building process. The numbers of neurons in input and output layers are determined by the numbers of input and output data, while the numbers of neurons in the hidden layers depend either on the complexity of the problem being handled, or the type of neuron activation function of the given layer, or the training algorithm, or the size of training data. The neuron activation function decides about the perceptron's properties. It is a non-linear relationship between the signal of total neuron excitement and its response. The behaviour of the neuron, and, consequently, of the whole neural network, depends therefore on the type of the activation function applied. The functions that are most frequently applied are the logistic function, or the hyperbolic tangent, or the linear function with saturation. Once we have determined the proper network structure, we can proceed to the training process.

The most common and reliable method of training that type of network is the backpropagation method. The method first calculates the errors occurring in the output layer on the basis of input and standard signals, followed by the same in reference to the preceding layers and so on until the input layer has been reached. The backward projected errors are multiplied by the same coefficients that have been used for the multiplication of the transmitted signals, although, at that time, the signals are sent from output to input. The initial values of weights should be close to zero. However, the training iteration (it is referred to as an epoch) generally assumes the values that are higher at the beginning of the training process.

The Conjugate Gradients algorithm that modifies weights once during the performance of one epoch is another method of training multilayer perceptrons. Besides, there is also the Quasi-Newton method of training neural networks which produces better results in many cases in comparison to the backpropagation method, although the former can display the tendency to develop convergence to local minima.

The selection of a proper training data set is one of the main problems we come across when applying neural networks. The data used in the network training process and the scope of such data should represent the whole population of data describing the given phenomenon or process. When collecting the training data, we must know that there is a relationship between known input and output values. The appearance of accidental values in those data, being considerably different than other data, can interfere with the training process. We should take into account the fact that the error determined for the training set becomes reduced in the training process.

The process of validation, intended to prevent network overfitting (or excessive matching the network with the training data) is also applied. In that process, a certain number of cases of the training data set are used to carry out the network training progress control (the validation set is not applied directly during the network training process).

In order to increase the credibility of the final network model, we separate the third set of data from the training set, i.e. the test set. All the essential cases should be represented in those three sets. The training process is concluded upon completion of either a certain number of epochs or upon reaching a specific error level. If the network is not able to attain the assumed error level, new neurons must be added to the hidden layer, or a whole neuron layer must be added. We can also remove a certain number of hidden neurons or a whole neuron layer when the network has been overfitted.

Application of neural networks to the prediction of specific gas pollutions: selection of the set of data used in the network training process

A considerable popularity of neural networks and wide areas of their application caused that simple software was developed allowing for neural network implementation. We used the Statistica software package in our studies for that purpose. Gas pollution forecasts were provided for SO_2, NO_2, NO at the stations of the Polish State Environmental Monitoring System in Zakopane and Nowy Sącz. In addition, ozone concentration forecast was done for Zakopane. The forecast of this parameter for Nowy Sącz is not possible, because it is not measured there. Neural networks representing various structures were built and tested for each type of gas pollution to select the best ones. Each network contained one layer of hidden neurons. The set of the data used in the training process was randomly divided into three subsets: training (70%), validation (15%), and testing (15%) subsets.

The data determining the concentrations of particular gases polluting the air and the meteorological data were included in the data set used in the network training process. The selection of the data to be included in the training set was preceded by a correlation analysis. Based on the analysis results, for each of the gas pollutant, the training data set included such parameters as the average daily pollutant rate on the forecast day and the day before, average, maximum and minimum temperatures, previous day temperatures, and snow cover thickness (for Zakopane only). As to ozone, the training data set also included sunshine, fog, and wind duration parameters. The data concerning the air quality were obtained from the Polish State Environmental Monitoring System. The data concerning the meteorological parameters were obtained from the archives of the Institute of Meteorology and Water Management. All the data concerned the period from 1 January 2017 to 31 December 2017.

RESULTS OF FORECASTING GAS POLLUTION OF AIR AND THEIR DISSCUSION

For each of the predicted pollutants, the best neural networks we developed had the multilayer perceptron (MLP) structure, although the networks differed in the number of neurons in particular layers. We also obtained various levels of prediction accuracy.

Our research results are presented in Table 1. The best value of the correlation indicators for the training set was obtained for sulphur dioxide in Zakopane

(0.98). Also the lowest value of the average forecast error, defined as the distance between the expected (real) value and the forecast value occurred for the same pollution, but in Nowy Sącz (1.7 µg). Since various pollutants were predicted, Table 1 also specifies the ranges of air pollution concentration, within which the respective pollutants occurred.

Table 1 The characteristics obtained for the best neural networks

Measurement station and pollutant	Correlation coefficient	Average difference between the expected value and the predicted one [µg]	STD of average difference between the expected value and the predicted one	Range of pollutant concentration (min-max) [µg]
Zakopane, SO_2	0.98	1.87	2.39	1.6-71.9
Zakopane, NO	0.84	5.40	7.90	0-109
Zakopane, NO_2	0.90	4.75	4.55	4-90
Zakopane, O_3	0.92	7.63	5.62	3-114
Nowy Sącz, SO_2	0.92	1.70	2.19	1.3-58.7
Nowy Sącz NO	0.82	6.76	7.71	0-90
Nowy Sącz NO_2	0.82	4.88	3.92	6-78

In addition to the average values and standard deviation (STD), we also carried out the calculation of the median value, as well as the lower and upper quartiles for the errors obtained in particular forecasts. The results of our analysis are presented in Fig. 1 below.

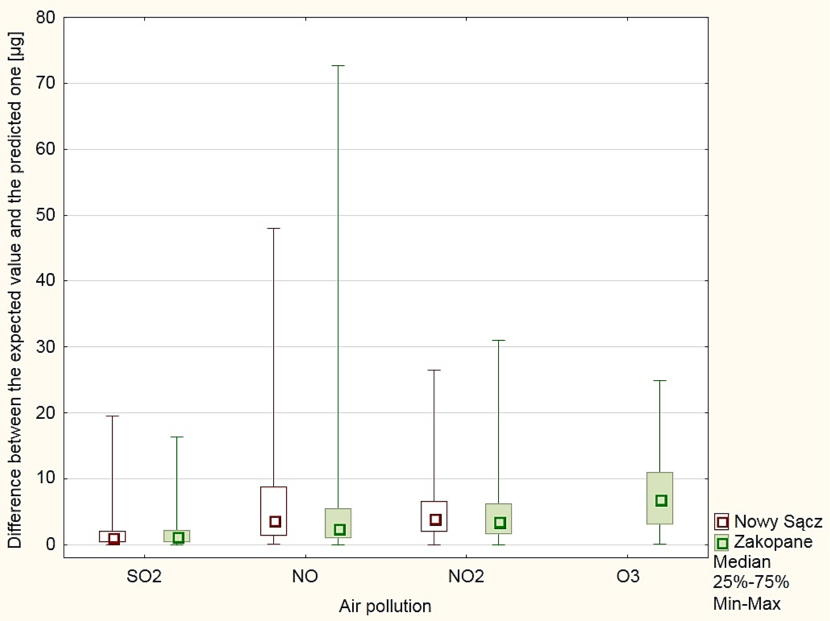


Fig. 1 Distribution of error values for particular pollutants

The lowest maximum error was obtained for sulphur dioxide. For the remaining pollutants, maximum errors reached high values (especially those of NO). However, when analysing the usability of the models developed by us, it was necessary to remark that, in the cases of SO_2, NO, and NO_2, 75% of the results obtained lied within the error boundary below 10 µg. Only for O_3, the value amounted to 10.95 µg.

The quality of the drafted forecasts is also presented in Figures 2-5 on the example of Zakopane.

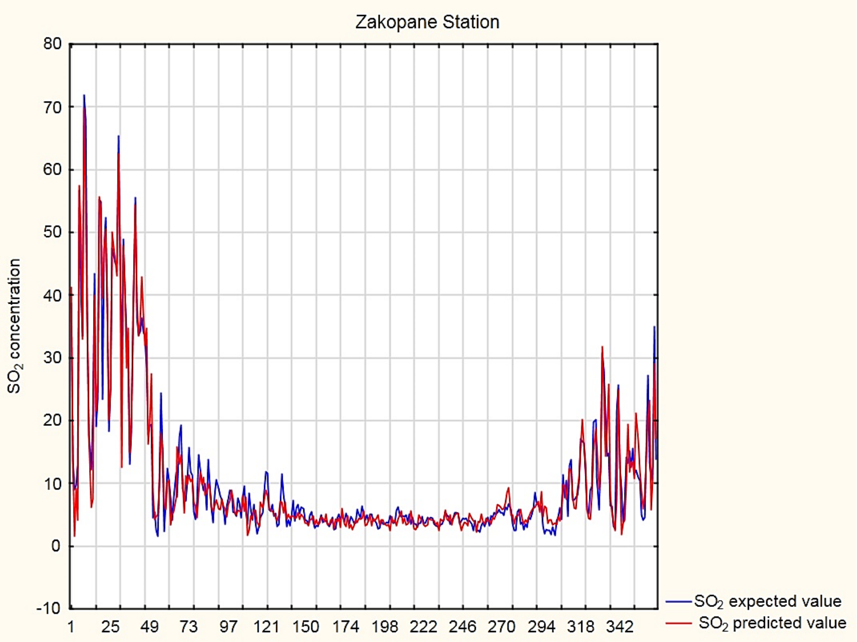


Fig. 2 Real and predicted values of SO_2 concentrations obtained from the MLP model for Zakopane Station

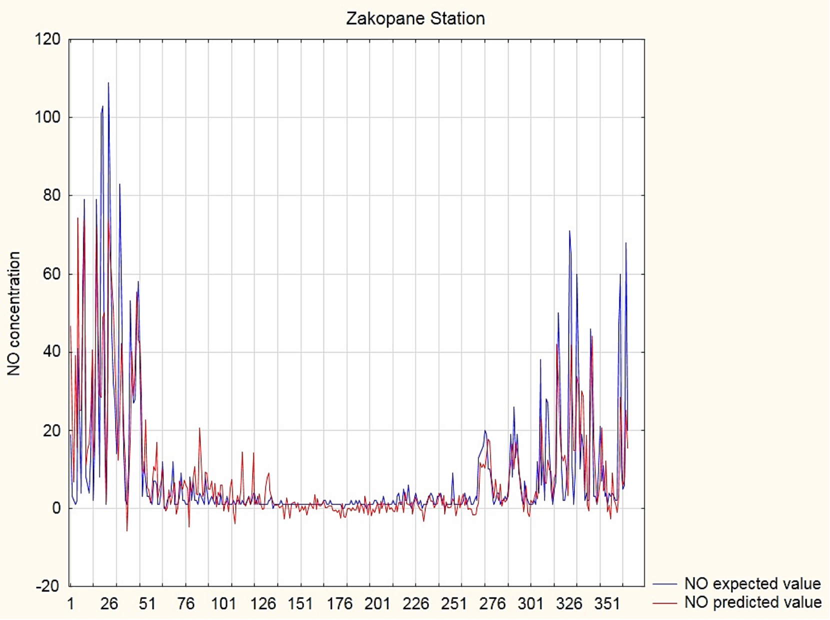


Fig. 3 Real and predicted values of NO concentrations obtained from the MLP model for Zakopane Station

The blue line represents the expected (real) value and the red line refers to the forecast value. An analysis of those graphs also displayed the fact that the best forecasts were obtained for sulphur dioxide.

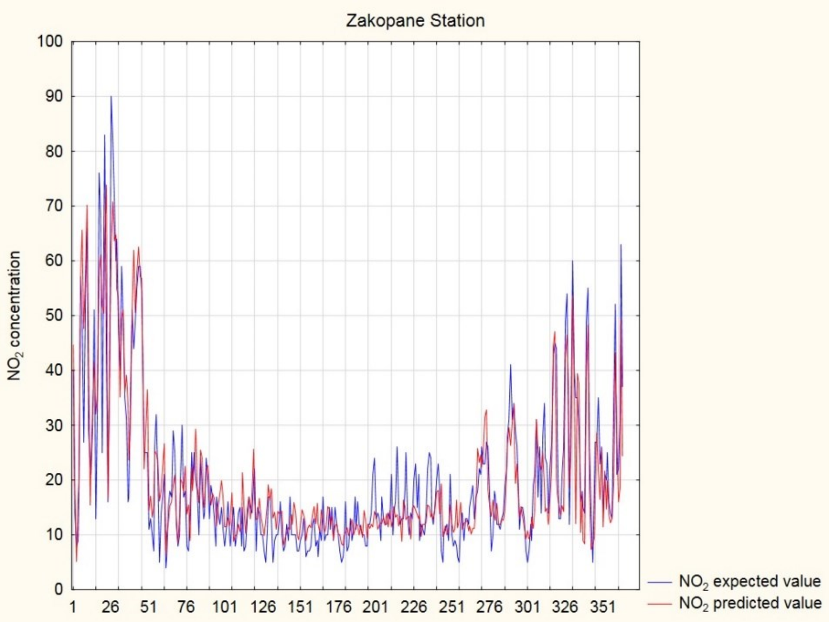


Fig. 4 Real and predicted values of NO2 concentrations obtained from the MLP model for Zakopane Station

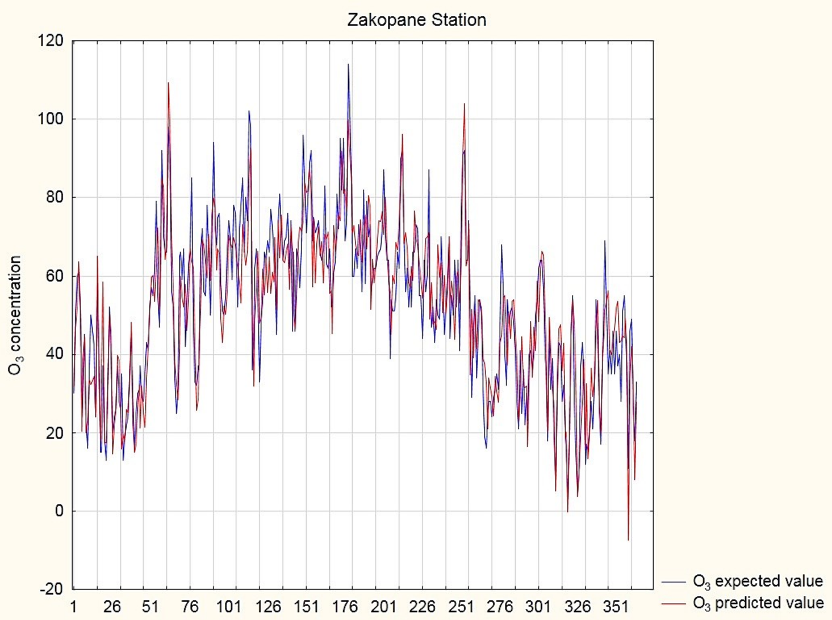


Fig. 5. Real and predicted values of O3 concentrations obtained from the MLP model for Zakopane Station

The accuracy of the models obtained differs. If we take into account the models built for Zakopane and Nowy Sącz, the input data are slightly different. This may cause differences in the quality of the neural networks. In addition, the local conditions for the spread of pollutants and their variability over time may influence on the quality of the model. On the other hand, the concentrations of particular pollutants in the air depend not only on meteorological data but also on other factors not included in the analyzed models, for example traffic road. These factors affect the concentrations of individual pollutants in varying degrees. The result may be, for example, differences in the accuracy of predictions of sulfur dioxide and nitrogen oxides. The nitrogen oxides are more dependent on traffic road and they are predicted worse than sulfur dioxide.

CONCLUSIONS

The issue of predicting pollution levels of air is fairly complex. Based on our research results presented here, one can conclude that we can create models of air pollution forecasts, using neural networks, without any reference to mathematical models representing the spread of pollution in the air.

However, it is necessary to train neural networks with a large data set, representing a broad range of measurements, in order to obtain high-quality forecasts. Such data should be properly selected. The input data should include those that essentially influence the concentrations of the pollutants that are objects of the forecasting process.

In the issue under discussion, we used meteorological data and pollutant concentration values. Consideration of additional factors would probably improve the quality of the models.

For each of the gas pollutants, the best results were obtained with the network representing a multilayer perceptron structure, with the Quasi-Newton algorithm. The correlation coefficients obtained for particular networks exceeded the value of 0.82.

A statistical analysis of forecast errors was also conducted to evaluate the quality of the network models. The best results were obtained for the SO_2 forecast. The average error rates for that pollutant was estimated at 1.87 µg for Zakopane and 1.7 µg for Nowy Sącz, with 75% of errors not exceeding 2.15 µg and 2.01 µg, respectively. The highest average error rate was obtained for ozone (7.63 µg). Moreover, the error boundary was higher for that pollutant: 75% of our results lied below that boundary (10.95 µg).

Based on the conducted analyzes, we can conclude that the neural networks allow to predict the level of gas air pollution concentration with satisfactory accuracy.

REFERENCES

Abderrahim, H., Chellali, M. R., and Hamou, A. (2016). Forecasting PM10 in Algiers: efficacy of multilayer perceptron networks. Environmental Science and Pollution Research, 23, pp. 1634-1641.

Hajto, M. J., Godłowska, J., Kaszowski, W. and Tomaszewska, A.M. (2012). System prognozowania rozprzestrzeniania zanieczyszczeń powietrza FAPPS – założenia, możliwości, rozwój. In: J. Konieczyński, ed. Ochrona powietrza w teorii i praktyce, 2, Instytut Podstaw Inżynierii Środowiska PAN, Zabrze, pp. 89-96.

Haupt, S. E., Pasini A. and Marzban C. (2009). Artificial Intelligence Methods in the Environmental Sciences. Springer.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R. and Cawley G. (2003). Extensive evaluation of neural network models for the prediction of NO_2 and PM 10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmospheric Environment, 37, pp. 4539-4550.

Kwiecień, J. and Pawul, M. (2012). Application of artificial neural networks to spring water quality prediction. Polish Journal of Environmental Studies, 21(5A), pp. 271-275.

Pawul, M. and Śliwka, M. (2016). Application of artificial neural networks for prediction of air pollution levels in environmental monitoring. Journal of Ecological Engineering 17, pp.190-196.

Rybarczyk, Y. and Zalakeviciute, R. (2018). Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review. Applied Sciences, 8, 2570.

smog.imgw.pl, (2019). IMGW Official Website. [online] Available at http://smog.imgw.pl/content/model [Accessed 10 Apr. 2019].

Tadeusiewicz, R. (1993). Neural Networks. Warszawa: Akademicka Oficyna Wydawnicza.

Tadeusiewicz, R. and Dobrowolski, J.W. (2004). Artificial intelligence and primary prevention of health hazards related to changes of elements in the environment. Polish Journal of Environmental Studies, 13(3), pp. 349-352.

www.who.int, (2019). WHO Official Website. [online] Available at www.who.int/airpollution/en/ [Accessed 15 Apr. 2019].

Abstract.
The issue of projecting the air pollution levels is quite essential from the viewpoint of the necessity to adopt specific prevention measures intended to reduce the pollution concentration in the air. One can apply certain machine learning methods, including neural networks, to build pollution concentration models. Neural networks are characterised by the fact that they can be used to solve the relevant problem when we face shortage of data, or we do not know the analytical relationship between input and output data. Consequently, neural networks can be applied in a number of problems. This paper discusses a possibility to apply neural networks to the prediction of selected gas concentrations in the air, based on the data originating from the measurement networks of the Polish State Environmental Monitoring System, combined with local meteorological data. Forecast results have been presented here for SO_2, NO, NO_2, and O_3 in various locations. The author also discusses the accuracy of the respective forecasts and indicates the relevant contributing factors.

Keywords: air quality monitoring, air pollution, artificial neural networks, prediction