Artificial Neural Network for Environmental Air Emissions Prediction Systems

DANUTA MATEI*, BOGDAN DOICIN, DIANA CURSARU, DORIN STANICA EZEANU
Petroleum-Gas University of Ploiesti, 39 Bucharest Blv., 100520, Ploiesti, Romania

Artificial neural network ANN is an appropriate tool for predestining the different relationships across many scientific researches. Air emissions prediction is an effective method of securing public health by revealing the imminent danger of air pollutants. So far, the existing methods of air emissions concentration prediction became useless on long-term dependencies, and most neglect spatial correlations. In this study, an artificial neural network model, which also takes into account the spatial-temporal variable, is proposed for prediction of air emission. The ANN was used in order to take into account the historical data, auxiliary data such as environmental quality and weather conditions. Evaluation of the results revealed that the developed ANN model appears to be an appropriate tool in predicting environmental air emissions. The data were collected from 12 air quality-monitoring stations from a random county in Romania. The proposed model performed well and presented a relative error varies from 0 to 4.7%.

Keywords: emissions, artificial neural network, air

At the beginning of an air emissions characterization procedure it is necessary to know the nature of pollutants, whether they are organic or inorganic nature, and the extent of pollution.

EPA (Environmental Protection Agency) has set national air quality standards for six common air pollutants (also called the criteria pollutants): nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), particulate matter (PM), carbon monoxide (CO), and lead (Pb). Four of these pollutants (CO, Pb, NO_2 and SO_2) are emitted directly from a variety of sources. Ozone is not directly emitted but is formed when oxides of nitrogen (NOx) and volatile organic compounds (VOCs) react in the presence of sunlight. PM can be emitted, or it can be formed when emissions of NOx, sulfur oxides (SOx), ammonia, organic compounds and other gases react in the atmosphere [1]. Uncontaminated air, emissions are often associated with important problems especially on human health due to the vast size of area contaminated to investigate and the large number of samples needed for analysis.

Energy related missions account for over two thirds of the greenhouse gases emissions. Global warming, as one of the most important issues in the recent century, is the cause of continuing rise in the average temperature of Earth's atmosphere and oceans level and is due to increased concentrations of greenhouse gases in the atmosphere caused by human activities especially because of burning of fossil fuels [2].

In recent years, many research attempts have centered on enriching approaches to predicting air emissions concentration. Usually methods of predicting air pollutant concentrations are separated in two main categories: deterministic and statistical methods. Deterministic methods are based on meteorological facts, while the statistical methods take into account the dispersion, transformation, diffusion and removal processes of pollutants based on chemical reactions and atmospheric physics. Statistical methods, however, avoid complicated theoretical models and simply implement based models to predict air quality. Widely used methods include so this called artificial neural network [3]. Multiple ANN structures have been developed [4-6] in order to reveal the secondary effects of different pollutant not only for air emissions but also regarding other process.

In recent years, the ANN appears to be the main tool in prediction of environmental factors pollution. In Hamadache et al. [7] the artificial neural network model was developed for predicting the oral acute toxicity on rats of a series of 77 herbicides based on their molecular structure, represented by 1666 calculated descriptors. The built ANN model was assessed comprehensively (internal and external validations). More than that, the robustness and predictive power of the model was verified by the different values. According to the obtained results, the ANN model developed and the mathematical formula obtained using the ANN approach for predicting the acute oral toxicity of herbicides to rats gave correct and acceptable results. In Pentos et al. [8] the artificial neural networks produced accurate mathematical models of relationships between soil texture, soil conditions and horizontal deformation as independent variables and traction force as well as traction efficiency as dependent variables. Based on neural models, the authors calculated the relative importance of explicative variables.

The aim of the present study was to find a fast, analytical procedure to predict the long term spatial temporal dependency of air emission concentration and a prediction framework which can forecast the air pollutant using ANN.

Experimental part
Materials and methods
Data description-case study area

For the very beginning due to technical reasons, we will note the random town chosen for air pollution emissions investigations as town CX. In downtown CX, air quality is monitored with the help of 6 automatic stations and 6 manual stations. The automatic stations are:

- CX-1 (APM headquarters) and CX-5 stations: they monitor the impact of traffic on the environment. The pollutants monitored are those specific to the transport activity, namely SO_2, NO, NO_2, NOx, CO, metals, PM10, Benzene, Toluene, o-Xylene, Ethylbenzene, m-p-Xylene.
- Stations CX-4 and CX-6 are stations that highlight the influence of industrial emissions on the pollution level. The

Station code	Station type	Pollutant	Valid date%
CX1	Traffic	NO₂, SO₂, CO, C₆H₆ PM10, Pb, Cd, Ni, As	30, 98.8, 0, 97.6, 97.5, 98.6, 97.5, 98.6, 97.5
CX2	Urban	NO₂, SO₂, CO, O₃, C₆H₆ PM 2,5	0, 88.6, 0, 45.8, 91.2, 93.1
CX3	Suburban	NO₂, SO₂, CO, O₃, C₆H₆ PM10, Pb, Cd, Ni, As	93.8, 97.5, 0, 94.9, 11.4, 95.8, 95.8, 94.9, 95.8, 94.7
CX4	Industrial	NO₂, SO₂, CO, O₃, C₆H₆ PM10	0, 70, 0, 93.2, 96.4, 30.1
CX5	Traffic	NO₂, SO₂, CO, C₆H₆ PM10, Pb, Cd, Ni, As	70.1, 91.2, 0, 78.4, 78, 78, 77, 78, 78
CX6	Industrial	NO₂, SO₂, CO, O₃ PM10, Pb, Cd, Ni, As	98.6, 94.8, 89.1, 0, 83.2, 82.1, 83.2, 83.2, 84.3

Table 1
DATA PROVIDED BY STATIONS IN CX COUNTY

monitored pollutants are SO_2, NO, NO_2, NOx, CO, O_3, PM10, Benzene, Toluene, Xylene, Ethylbenzene, m-p-Xylene (CX4), SO_2, NO_2, NOx, CO, O_3, metals, PM10 (CX-6).

- Station CX-2 urban background station was located in residential area, away from local emission sources. Pollutants monitored SO_2, NO, NO_2, NOx, CO, O_3, PM10, metals, Benzene, Toluene, o-Xylene, Ethylbenzene, m-p-Xylene.

- CX-3 station Suburban Substation Station, assesses the influence of the *following settlements* on air quality. Pollutants monitored SO_2, NO, NO_2, NOx, CO, O_3, PM10, metals, Benzene, Toluene, o-Xylene, Ethylbenzene, m-p-Xylene. All the data provided by stations in CX county are presented in table 1.

The 6 manual stations are located as follows:

- Two stations are located in the Eastern area of town: which monitors the activity of the units located in these area especially crude oil refineries. Thus, the following indicators are taken and analyzed: H_2S, formaldehyde (HCHO), NH_3, suspension powders.

-The station, which monitors the activity of some commercial factories and road traffic and the railway station in the zone, which monitors the activity on the northern platform of the city. Thus, the following indicators are taken and analyzed: NH_3, HCHO, H_2S, phenol, suspension powders.

-The station located in the west part of the city that monitors economic activity in the area and road traffic, the following indicators are taken and analyzed: H_2S, HCHO, NH_3, suspended powders.

-The station that monitored the city center, a point chosen to track the dispersion of pollutants due to industry and especially the three oil refining refineries surrounding the city. The indicators taken and analyzed are: NH_3, H_2S, HCHO.

The air quality indicators used for ANN model were collected for 12 months and are depicted in table 2.

ANN model

In order to perform an accurate estimation of the pollutants, mathematical modeling using artificial neural networks was used. An ANN is an algorithm, which simulates the functioning of animal nervous central system, especially the brain. Once trained, an ANN can predict at a high speed the information it was designed to predict to begin with [9].

To maintain the similarity with the biological nervous systems, an ANN is built from many computational devices, called *neurons*. The neuron of an artificial neural network has the structure presented in Figure 1 [10]:

From Figure 1 it can be noticed that a neuron works based on some mathematical calculations. Thus, the input (noted with p) is multiplied by a weight, noted with w. The result of this multiplication is summed with the bias, noted with b. The resulting value is noted with n and it is called net input. The net input becomes the argument for the function f, which is called transfer function (or activation function). The result returned by the transfer function is noted with a and is the neuron output.

For a given set of values for w, p and b, the value of the output depends of the transfer function that is utilized. A couple of the most utilized transfer functions are presented in Table 3 [10].

The structure of the neuron from Figure 1 is presented for the case in which the neuron receives only one input. This was made intentionally, for simplicity reasons. In practice, a neuron very often receives two or more inputs. The only difference in this case is that the input, the weight and the bias are not scalars, but matrices and the subsequent computations are made considering this fact. Each neuron from a layer have the same structure presented in Figure 1. The structure of a layer of neurons is presented in Figure 2 [10]:

To successfully train a network, using supervised training, a *training database* is needed. A *training database* is a collection of input-output pairs, which tell the network
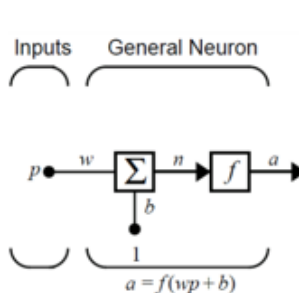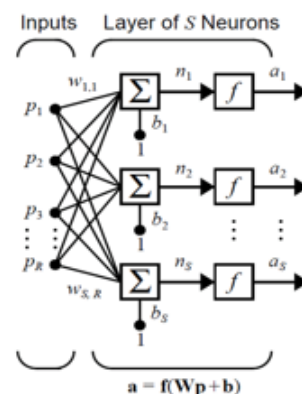

Fig. 1. The structure of a neuron from an ANN

Fig. 2. Structure of a layer of neurons

Table 2
CENTRALIZER ON AIR QUALITY INDICATORS JANUARY - DECEMBER 2017

Station	Pollutant	Number of determination	Number of determination exceeded	% determination exceeded	Maximum admissible concentration (mg/m^3)	Average concentration (mg/m^3)
West part of the city	HCHO	200	0	0	0.011	0.0028
	NH₃	200	0	0	0.100	0.0301
	H₂S	200	0	0	0.008	0.0029
	Suspensions	200	5	2.5	0.140	0.1471
Nord part of the city I	HCHO	200	28	10.3	0.012	0.0071
	NH₃	200	14	5.6	0.100	0.0424
	H₂S	200	2	0.5	0.009	0.0031
	Suspensions	200	17	8.3	0.150	0.1462
Eastern part of the city I	HCHO	200	1	0.4	0.012	0.0037
	NH₃	200	0	0	0.101	0.0288
	H₂S	200	2	0.5	0.008	0.0028
	Suspensions	200	0	0	0.155	0.1259
Central zone	HCHO	199	19	5.62	0.012	0.0054
	NH₃	199	15	7.04	0.100	0.0480
	H₂S	199	1	0.8	0.009	0.0033
Eastern part of the city II	HCHO	200	13	4.7	0.012	0.0063
	NH₃	200	0	0	0.104	0.0305
	H₂S	200	3	0.8	0.008	0.0033
Nord part of the city II	HCHO	200	12	6.6	0.013	0.0078
	phenol	200	2	0.8	0.03	0.0085
	H₂S	200	1	1.3	0.009	0.0033

the proper behavior. The general structure of these pairs is presented in equation (1).

$$\{p_1, t_1\}, \{p_2, t_2\}, \dots, \{p_n, t_n\} \qquad (1)$$

In equation (1), the values p_i, i=1...n represent inputs to the network and the values t_i, i=1...n represent the correct corresponding outputs (target outputs). As the inputs are applied to the network, by training, the network outputs are compared to the targets.

The artificial neural network used in this paper will be a network with a single hidden layer, trained using a supervised training algorithm.

Network framework

The neural network in this paper was created and configured using Matlab R2017b. Among the pollutants studied and presented in Table 1, NH_3 and H_2S are, by far, the most important and dangerous pollutants from the county. Therefore, the artificial neural network will be trained to predict the concentration of these two pollutants.

The stations mentioned in Table 1 have an automatic method of sampling, so more samples are taken in 24 hours. In this paper, 10 samples will be used, from each station, to build the training database for the artificial neural network. Thus, it will be able not only to predict the pollutant concentration, but to predict the pollutant concentration at a specific hour. The usefulness of this prediction is obvious when we take into account the fact that throughout the day many other outside factors influence the measured values: traffic variation, factory production variation, etc.

Having considering all this information, the artificial neural network that will be used has the station number and the day period as inputs. The outputs are the NH_3 concentration and the H_2S concentration. The day period is quantified as the sample number, as it is impractical to enter it into the database in the raw form.

A sample of the used database is presented in Table 4.

Table 3
THE MOST UTILIZED TRANSFER FUNCTIONS

Name	Mathematical Relation
Hard Limit	$f(n) = \begin{cases} 0, n < 0 \\ 1, n \geq 0 \end{cases}$
Symmetrical Hard Limit	$f(n) = \begin{cases} -1, n < 0 \\ 1, n \geq 0 \end{cases}$
Linear	$f(n) = n$
Saturating Linear	$f(n) = \begin{cases} 0, n < 0 \\ n, n \in [0,1] \\ 1, n > 1 \end{cases}$
Symmetric Saturating Linear	$f(n) = \begin{cases} -1, n < -1 \\ n, n \in [-1,1] \\ 1, n > 1 \end{cases}$
Hyperbolic Tangent Sigmoid	$f(n) = \dfrac{e^n - e^{-n}}{e^n + e^{-n}}$
Log-Sigmoid	$f(n) = \dfrac{1}{1 + e^{-n}}$

Table 4
SAMPLE TRAINING DATABASE

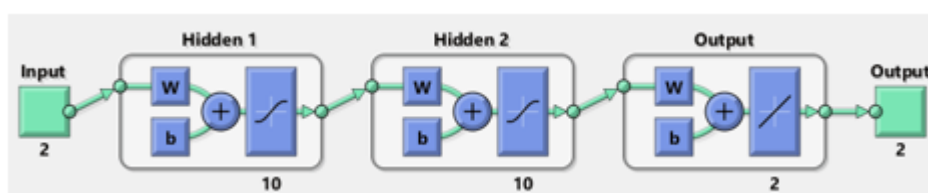| Station number	Sample number	NH_3 concentration	H_2S concentration
1	1	0.0235	0.0076
1	2	0.0326	0.0085
1	3	0.0258	0.0099
1	4	0.0232	0.0073
1	5	0.048	0.0086
1	6	0.1	0.0092
1	7	0.0728	0.0075

Fig. 3. Structure of the artificial neural network

In Table 4, and in the entire database, the station number is coded exactly as it is in Table 1, removing the *CX* prefix. Thus, station number 1 is the station CX 1, from the first table.

The entire training database has data from 5 stations and for each station 10 samples were taken. Thus, the entire database has 50 samples.

The next step in establishing the artificial network framework was to determine the number of neurons that the hidden layer will have. Because there is not a clear algorithm of how to determine this number, the trial and error method was used. This method showed that the configuration, which gives the best approximation, has two hidden layers, each one with 10 neurons. Increasing the number of neurons for any of the two layers proved to be overkill, leading to over fitting.

Among the entire training algorithms Matlab offered, the Bayesian regularization back propagation algorithm was chosen to train the artificial neural network. The reason is that it offered the most accurate prediction. Its main drawback is an increased length of time needed for training, but the results justify the waiting.

Thus, the structure of the artificial neural network is presented in Figure 3.

In Figure 3 it can be noticed the neuron structure that was presented earlier. Each neuron from the two hidden layers have their own weights and biases, the transfer function being the hyperbolic tangent sigmoid for both of these layers (presented in Table 3).

Results and discussions

Before the artificial neural network can be used to predict the concentration of the two studied pollutants, it is necessary to assess the training efficiency. The higher the training efficiency is, the more precise the predictions will be.

For this experiment, two criterions were chosen: the regression analysis and the error histogram.

The regression analysis has the purpose of measuring how correlated the data from the training database are. The output of the analysis is a number, between 0 and 1, noted with R. If R equals 1, it means that the data is perfectly correlated. If R equals 0, it means total randomness between the data.

The regression analysis was performed for each of the categories the data were fit into and for the database as a whole. The result is presented in Figure 4.

In Figure 4 it can be seen that training database is very highly correlated, R having values between 0.97 and 0.99. This means that the ANN predictions will be accurate, with respect to the database provided for training.

The error histogram is another way of rating the training efficiency. Instead of focusing on a single value, the error histogram shows more in detail how the training went, by showing the differences between target data and training data. The more of these differences are close to 0, the higher the training efficiency is.

The error histogram results are presented in Figure 5.

The results from Figure 5 confirms what the regression analysis from Figure 4 showed: the training was efficient. In Figure 5, most of the errors were between -0.00095 and 0.00056, which is very close to 0.

Because the two chosen criteria clearly showed highly training efficiency, the authors consider that studying other training efficiency criteria is useless, because they will lead to the same result.

Prediction performance

To assess the prediction performance, the simplest and most efficient way is to compare the measured results for the two studied pollutants with the predicted results for the same inputs. However, because the data from the training database are distributed randomly in the three categories, a random factor will always intervene.

In this experiment, because the data from the training database are highly correlated (as it was proved in Figure 4 and Figure 5), the authors considered to make 10
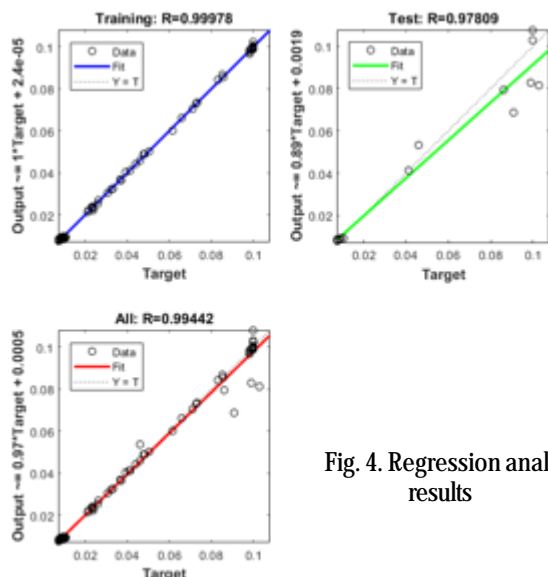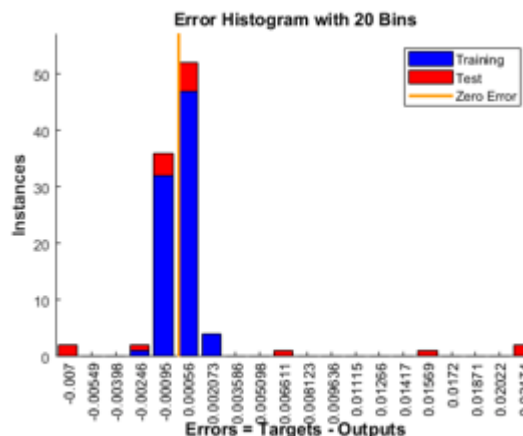
Fig. 4. Regression analysis results

Fig. 5. Error histogram results

Measured concentrations		Determined concentrations		Relative error (%)	
NH_3	H_2S	NH_3	H_2S	NH_3	H_2S
0.0235	0.0076	0.0235	0.0076	0	0
0.0326	0.0085	0.0326	0.0085	0	0
0.0258	0.0099	0.0258	0.0099	0	0
0.0232	0.0073	0.0232	0.0074	0	1.36
0.048	0.0086	0.048	0.0085	0	1.17
0.01	0.0092	0.01	0.0092	0	0
0.0728	0.0075	0.0728	0.0075	0	0

Table 5
PREDICTION PERFORMANCE FOR THE ARTIFICIAL NEURAL NETWORK

estimations for each input and to use the average of the obtained values.

The inputs considered to determine the prediction performance are those presented in Table 4. The obtained results are presented in Table 5.

As it can be seen in Table 5, the ANN offered very accurate predictions of the NH_3 and H_2S pollutants from the stations taken into account. The relative errors vary from 0 to 1.36%. Based on these findings, the authors conclude that artificial intelligence and artificial neural networks can be successfully used to predict the pollutants concentration either in air or water.

The ANN that was used had two hidden layers, thing that lead to almost perfect predictions, taking into account the data from the training dataset. Having only one hidden layer lead to prediction errors up to 20%, unacceptable. An added bonus to using the aforementioned method is that the user can customize the neural network according to his needs. It is not mandatory to have only two pollutants monitored, as an example. The method presented in this paper can successfully be applied in other cases as well.

Conclusions

This study presents an ANN model for predicting air pollutant concentration based on previous data registered. The ANN model is able of modeling series with variable dependence and can automatically offer the optimum times. To improve the performance of the actual model data regarding main pollutant were collected from 12 air quality-monitoring stations. Experiments concluded that the proposed ANN is valuable for better prediction performances.

References
1. DONAHUE, N.M., Chapter 3.2 -Air Pollution and Air Quality, An Inclusive Approach,US National Library of MedicineNational Institutes of Health, 2018, p. 151–176;
2. KHOSHNEVISAN, B., OMID M., Prognostication of Environmental Indices in Potato Production Using Artificial Neural Networks, Journal of Cleaner Production,**52,** 2013, p. 402-409;
3. LI X., PENG L., YAO X., CUI S.,HU Y.,YOU C.,CHI T., Long Short-term Memory Neural Network for Air Pollutant Concentration Predictions: Method Development and Evaluation,Environmental Pollution, **231**, 2017, p.997-1004;
4. BONELLI, M.G., FERRINI M, MANNI A.,Artificial Neural Networks to Evaluate Organic and Inorganic Contamination in Agricultural Soils, Chemosphere, **186**, 2017, p.124-131;
5. MATEI, D., DOICIN B., CURSARU D.L., Pd/SBA-15 Mesoporous Catalyst for Ethanol Steam Reforming. A Neural Network Approach, Digest Journal of Nanomaterials and Biostructures, **2,** 2016,p. 443-451;
6. DOICIN, B., EZEANU, D.S., CURSARU, D., Estimation Relationships for Ethanol Conversion and Hydrogen Yield Using Neural Network for Ethanol Steam Reforming on Co/Al_2O_3 Catalyst, Rev. Chim. (Bucharest), **68,** no.11, 2017, p. 2720-2725;
7. HAMADACHE, M.,HANINI S., BENKORTBI O., AMRANE A., KHAOUANE L.,MOUSSA C.S., Artificial Neural Network-based Equation to Predict the Toxicity of Herbicides on Rats, Chemometrics and Intelligent Laboratory Systems,**154**, 2016, p. 7-15;
8. PENTO, K., PIECZARKA K., Applying an Artificial Neural Network Approach to the Analysis of Tractive Properties in Changing Soil Conditions, Soil & Tillage Research, **165**, 2017, p.113-120;
9. SHIVAKUMAR A., SHRINIVR.B.R., SAMAGA B.S., Performance and Emission Characteristics of a 4 Stroke CI Engine Operated on Honge Methyl Ester Using Artificial Neural Network, Journal of Engineering and Applied Sciences, **5,** 2010, p. 83-94;
10. HAGAN, M., DEMUTHH.B. , BEALE M.H. ,Neural Network Design, 2nd Edition, Oklahoma State University Publishing House, 2010, p. 665-670.