Technical Sciences
Academy of Romania
www.jesi.astr.ro

Journal of Engineering Sciences and Innovation

Volume 5, Issue 2 / 2020, pp. 141-148

http://doi.org/10.56958/jesi.2020.5.2.141

C. Environmental Engineering and Energy

Received ***10 March 2020*** ***Accepted 19 June 2020***
Received in revised from ***21 May 2020***

Forecast of the evolution of a pollutant concentration from a groundwater source

ALEXANDRU WOINAROSCHY[1*], ALICE IORDACHE[2]

[1]Technical Sciences Academy of Romania, 26 Bul. Dacia, 030167 Bucharest, Romania
[2]Department of Chemical and Biochemical Engineering,
University POLITEHNICA of Bucharest, 1-8 Polizu Str., 011061 Bucharest, Romania

Abstract. The case study presents the forecast of the concentration of nitrate in drinking water from the Mangalia reservoir. The historical database includes monthly measurements of the respective concentration for the period January 2013 - October 2019. The forecast of the evolution of nitrate concentration was made with a multilayer neural network (MNN) with 16 input neurons, 10 hidden neurons, and 4 output neurons. The neural network was trained using back propagation algorithm with dynamic "window" training data set. The predictions for the year 2019 of the trained network have a maximum error of 21.02%, the average error being 4.89%. The case study shows that the use of MNNs is a promising way for forecasts in the field of drinking water quality management.

Keywords: forecast, nitrate concentration, drinking water, neural network

1. Introduction

Time series analysis with explanatory variables encompasses methods to model and predict correlated data [1]. The various techniques in this field have the mutual goal of reproducing the output series with reliability and accuracy from the estimation of the input series. Time series techniques can essentially be divided into two sets of methods: univariate and multivariate. In the case of univariate methods the output series is explained by a constant portion and/or trend, seasonality, and in many cases, by the series lagged in time. Multivariate methods use the influence of other variables on the behaviour of the output series to obtain better results in the representation of a transfer function. One of the main areas of application of such methods is in the environmental science. Nunnari et al. [2]

[*] Correspondence address: a_woinaroschy@chim.upb.ro

compared several statistical techniques for modelling SO_2 concentration using the information of wind direction, wind speed, solar radiation, temperature and relative humidity. Andriyas and McKee [3] used biophysical conditions in farmers' fields and the irrigation delivery system during the growing season to anticipate irrigation water. Lima et al. [4] developed a forecasting model for the water inflow incorporating the effect of climate variables like precipitation. Sfetsos and Coonick [5] compared the approaches on predicting solar radiation, and Porporato and Ridolfi [6] forecasted river flows.

It were established [1] that in 2018 the field of environmental sciences had as the number of published articles the first weight (12%) of the total 10 fields of application

of time series modelling. It is also evident the increase in the number of forecast applications, from 49 in the period 1967-1998, to 142 in the period 2013-2016.

The two main approaches to these applications are regression models and artificial neural networks. A comparison of regression models with artificial neural networks (ANN) is as follows [1]:

Regression models:

Strengths:

- Does not require high computational power;
- The relationship between the exogenous and response variables is open and most of the time, interpretable.

Weaknesses:

- Is sensitive to outliers;
- It presupposes that the model errors are independent and follow a normal distribution with zero mean and constant variance.

ANN

Strengths:

- High data processing power due to its massively distributed structure and its ability to learn and therefore to generalise, producing suitable outputs for inputs that were not present during the training.

Weaknesses:

- Requires a high computational power and a large amount of data in the training process;
- Is considered a black-box model, since the relationship between exogenous and response variables is impossible to be interpret.ANN

Strengths:

- High data processing power due to its massively distributed structure and its ability to learn and therefore to generalise, producing suitable outputs for inputs that were not present during the training.

Two types of artificial neural networks [7] have been involved in the field of environmental sciences:

- Multilayer neural networks;
- Radial basis neural networks.

The performances of the two types of networks are comparable. The implementation of multilayer networks is easier.

2. Forecast of nitrate concentration from Mangalia groundwater source

The possibility of predicting the evolution of the concentration of a pollutant in a source of drinking water is particularly important: if this indicates the future increase of that concentration above the legally allowed limits, it is necessary to take appropriate measures to prevent this unwanted phenomenon. The respective study requires a historical database, consisting of a series of measurements of the concentration of the pollutant at different time intervals in a given period. The forecast of the concentration of nitrate in drinking water from the Mangalia (a city in SE Romania) reservoir is exposed as an application. The historical database includes monthly measurements of the respective concentration for the period January 2013 - October 2019, being presented in Table 1. The maximum allowed limit (Romanian Law 458/2002 and Romanian Law 311/2004) of the nitrate concentration is 50 mg/L. It should be noted that it was exceeded in January, July, September and October 2015 and in January - April 2016.

Forecast based on regression models

The best correlation using a regression model was obtained based on a tenth order polynomial and is represented in the Fig. 1. It can be seen from Fig. 1 that the correlation cannot be considered satisfactory for use in a forecast study.

Forecast based on multilayer neural networks

Neural networks have the ability to offer some human characteristics in solving problems that are difficult to solve using logical, analytical and software methods. These can be defined as a computer system consisting of a number of simple but strongly interconnected processing elements, which process information through its dynamic response. A multilayer neural network (MNN) is one of the simplest and the most well-known types of ANNs. The structure of MNN which includes input layer, hidden layer and output layer has a significant impact on predictive capability. Independent and dependent parameters determine the numbers of neurons in the input and output layer, respectively. Neurons numbers in the hidden layer are generally determined via the procedure using the trial and error observation. Insufficient hidden layer neurons make the network difficult to fully learn the data laws, causing under-fitting problems. However, excessive neurons may lead to over-fitting as a result of extra degrees of freedom [8, 9].

Table 1. The historical database

Month,	Nitrate concentration (mg/L)						
Year	2013	2014	2015	2016	2017	2018	2019
January	31.80	40.20	50.50	54.80	43.0	37.40	38.50
February	58.80	42.00	43.10	74.50	37.10	34.10	39.80
March	51.80	44.00	41.20	72.70	43.50	34.60	38.40
April	51.40	45.10	40.80	82.0	47.10	38.30	39.70
May	60.90	47.50	39.40	37.50	35.80	35.80	39.00

Month,	Nitrate concentration (mg/L)						
Year	2013	2014	2015	2016	2017	2018	2019
June	60.00	48.20	40.60	36.10	39.40	40.10	43.50
July	61.60	46.80	52.30	35.80	40.10	46.60	45.40
August	61.90	66.70	66.60	38.00	46.30	42.30	47.20
September	45.10	38.40	54.30	36.20	42.30	40.00	42.00
October	41.80	42.00	51.80	31.00	41.02	44.20	47.40
November	38.40	44.50	48.40	34.60	39.80	42.60	
December	37.90	44.30	52.40	31.10	37.60	40.20	

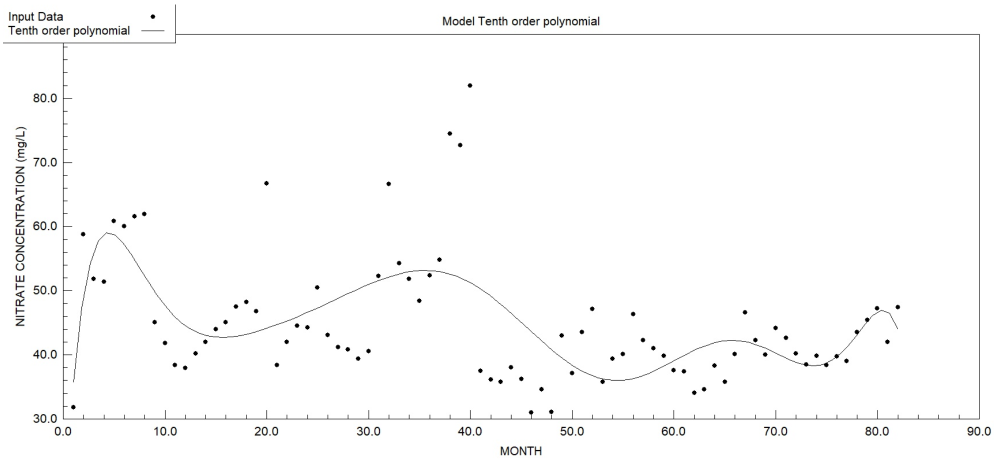

Fig. 1. Regression model using a tenth order polynomial.

In Fig. 2 is represented a MNN with 3 input neurons and bias, 5 hidden neurons, and 2 output neurons.

Based on specific neurons and weights contained in each layer, as well as an appropriate training algorithm, a well-trained MNN is capable of generating principle to track nonlinear input-output relationships, giving predicted values of the corresponding output(s) based on input conditions. A MNN trained by back-propagation (BP) algorithms is the most representative type of ANN which is widely applied in the field of environmental pollution controls [7]. Fig. 3 shows that the input signals are processed through the hidden layers, and the output signals are generated through nonlinear transformation.

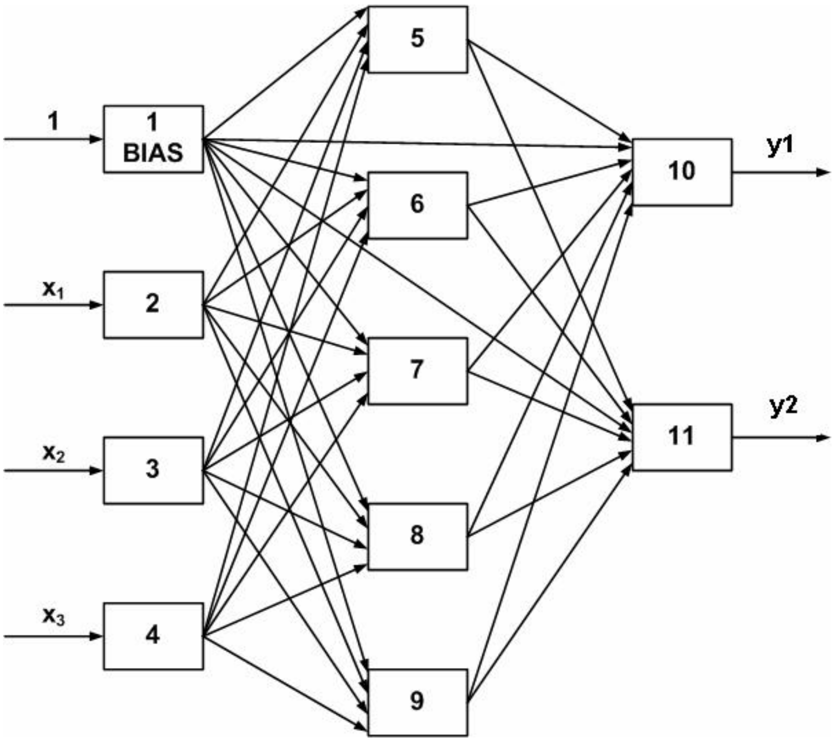

Fig. 2. A MNN with 3 input neurons and bias, 5 hidden neurons, and 2 output neurons.

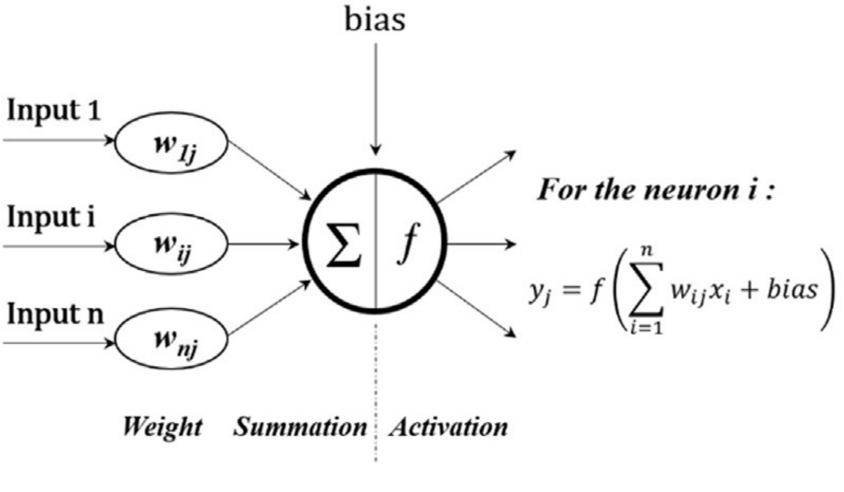

Fig. 3. Information processing in an artificial neuron.

The state of neurons in each layer only affects the neurons in the next layer. The correlation between the input x_i and output y_j from a neuron in the hidden layer can be expressed as follows:

$$y_j = f\left(\sum_{i=1}^{n} w_{ij}\, x_i + b\right) \tag{1}$$

where y_j is the *j*th output in the hidden layer, f(**x**) is the transfer function, n is the number of input variables, w_{ij} denotes the weight from element i in input layer to element j in the hidden layer, x_i is the ith output from the input layer, and b is the bias of hidden layer. The transfer function allows defining the state of the neuron based on inputs, it can be binary with limit (threshold) or linear with one threshold or more:

$$f(x) = \begin{cases} p, & x < p \\ x, & x \in [\,p, v\,] \\ v, & x > v \end{cases} \tag{2}$$

sigmoid with horizontal asymptotes:

$$f(x) = a \cdot \frac{e^{kx} - 1}{e^{kx} + 1} \tag{3}$$

or stochastic type:

$$f(x) = \frac{1}{1 + \exp\left(-\frac{x}{T}\right)} \tag{4}$$

The signals generated from the output neuron are the conversion of the weighted sum of output signals in the hidden layer. Standard BP algorithm is based on a gradient search, in which the network weights and thresholds move backwards along the performance function gradient, minimizing the errors between the actual output values of the network and the expected output values [9].

The training algorithm employed for weights correction can be expressed as follows:

$$\Delta w_{ij}\,(s+1) = -\,\eta \frac{\partial E}{\partial w_{ij}} + \mu \Delta w_{ij}\,(s) \tag{5}$$

where $\Delta w_{ij}(s)$ is expressed as the correction of the weight at the *s*th training step, η denotes the training rate, E denotes the total sum squared error of all data in the training set and μ is the momentum factor. The weights and thresholds of all the neurons are being updated until all the errors are located within the required tolerance or the maximum number of iterations is achieved.

The forecast of the evolution of nitrate concentration was made with a neural network with 16 input neurons, 10 hidden neurons, and 4 output neurons. The sigmoid activation function was used:

$$f(x) = \frac{1}{1+e^{-x}} \tag{6}$$

The BP algorithm has a training rate of 0.25 and a momentum factor of 0.5. The maximum number of iterations was 1,000,000, and the training limit error 0.01.

The training data set corresponds to a dynamic "window" [10]. The centre of the window corresponds to the reference time t0. As inputs to the network are 8 previous values and 8 future values (related to the reference time t_0) of the nitrate concentration. The desired answers correspond to the following 4 values of the nitrate concentration, after the 8 future values. It should be noted that because the database is "historical", future values (relative to the centre of the window) are known when are used in the training algorithm. The dynamic window scans the data set with a step Δt of one month. When the end of the database is reached, its exploration is resumed from the beginning, until the maximum number of training iterations are exhausted, or the limit value of error is reached.

At the end of 1,000,000 training iterations, the error is 0.0209. The predictions for the year 2019 of the trained network were analysed. The respective results are presented in Table 2.

Table 2. Measured and predicted values of nitrate concentrations

Month, Year	Measured nitrate conc. (mg/L)	Predicted nitrate conc. (mg/L)	Error %
January 2019	38.5	42	9.09
February 2019	39.8	39.6	-0.5
March 2019	38.4	43.5	13.28
April 2019	39.7	45.4	14.35
May 2019	39	47.2	21.02
June 2019	43.5	42.2	-2.99
July 2019	45.4	43.3	-4.63
August 2019	47.2	45.6	-3.39
September 2019	42	43.3	3.09
October 2019	47.4	47.2	-0.42

3. Conclusions

The errors between the measured and predicted values are reasonable, the maximum error being for May 2019 (21.02%), and the average error is 4.89%. Both the measured and predicted values are below the legal limit of 50 mg/L, so no measures were needed to limit the increase of nitrate concentration. This case study shows that the use of MNNs is a promising way for forecasts in the field of drinking water quality management.

References

[1]. Maçaira P. M., Tavares Thomé A. M., Oliveira F. L., Carvalho Ferrer A. L., *Time series analysis with explanatory variables: A systematic literature review*, Environ. Model. Software, **107**, 2018, p. 199-209.
[2]. Nunnari G., Dorling S., Schlink U., Cawley G., Foxall R., Chatterton T., *Modelling SO2 concentration at a point with statistical approaches*, Environ. Model. Software, **19**, 2004, p. 887–905.
[3]. S. Andriyas, M. McKee, *Recursive partitioning techniques for modeling irrigation Behavior*, Environ. Model. Software, **47**, 2013, p. 207–217.
[4]. Lima L. M. M., Popova E., Damien P., *Modeling and forecasting of Brazilian reservoir inflows via dynamic linear models*, Int. J. Forecast., **30**, 2014, p. 464–476.
[5]. Sfetsos A., Coonick A. H., *Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques*, Sol. Energy, **68**, 2000, p. 169–178.
[6]. Porporato A., Ridolfi L., *Multivariate nonlinear prediction of river flows*, J. Hydrol, **248**, 2001, p. 109–122.
[7]. Ye Z., Jiaqian Y. J., Zhong N., Tu X., Jia J., Wang J., *Tackling environmental challenges in pollution controls using artificial intelligence: A review*, Science of the Total Environment, **699**, 134279, 2020.
[8]. Wasserman P. D., *Neural Computing: Theory and Practice*, Van Nostrand-Reinhold, 1989.
[9]. Bulsari A. B., *Neural Networks for Chemical Engineers*, Elsevier, 1995.
[10]. Refenes A. N., Azema-Barac M., Chen I., Karoussos S.A., *Currency exchange rate prediction and neural network design strategies*, Neural Computing & Applications, **1**, 1993, p.46-58.