ORIGINAL PAPER



Simulation, evaluation and prediction modeling of river water quality properties (case study: Ireland Rivers)

E. S. Salami¹ \cdot M. Ehteshami²

Received: 24 May 2014/Revised: 28 January 2015/Accepted: 15 March 2015/Published online: 9 April 2015 © Islamic Azad University (IAU) 2015

Abstract In this analysis, three input parameters temperature, pH and electrical conductivity were chosen due to their easy and less costly measurement technique, and a package of six models were presented for estimating the concentrations of dissolved oxygen, DO percentage, biological oxygen demand, chloride, alkalinity and total hardness. 3001 data sets (a 3001×8 data array) were used to training the models. The models have been tested in order to verify their prediction values, and the resulted R factor (the rate of precision) for each model equals to 0.93, 0.95, 0.77, 0.82, 0.85 and 0.92, respectively. This proves that the package can be used to estimate the concentrations of water quality parameters with accuracy close to the reality. The River data collected from 210 monitoring stations located in all over Ireland have been used. The data set covers different conditions and makes the model applicable in many different places and conditions. For development of all models, feed-forward algorithm used for training, as well as the Levenberg-Marquardt and tansign(x) functions as learning and transfer functions.

Keywords Artificial neural networks · Ireland Rivers · Modeling · Water characteristics · Water quality

M. Ehteshami Maehtesh@gmail.com

¹ KN Toosi University of Technology, Tehran, Iran

Introduction

The situation of water resources in the world becomes more challenging every year. During the last century, water consumption grew at twice the rate of population increase (Stockholm International Water Institute and Elsevier 2012). In addition, the complexity of managing natural resources generally increases in parallel with human population growth (Varnell et al. 2008). Assessment of properties and processes related to running waters is a major issue in the management of aquatic environments (Schleiter et al. 1999). As a result, accurate determination of the concentration of nutrients and other substances in water bodies is an essential requirement for supporting effective management and legislation with regard to such environments (Donohue and Irvine 2008). Water management decisions are increasingly based on model studies (Scholten et al. 2007), while modeling tools are becoming progressively more sophisticated (McKnighta et al. 2010).

Modeling of water quality parameters also has many benefits, the most important of which are explained here. The value of data modeling can be demonstrated at the overall savings that it makes in terms of maintenance or development costs. When this issue is considered in broad terms across an organization's total budget, the amount of saving can be truly significant. The value of data modeling can be also seen at a more detailed level in terms of savings it provides through development tasks related to a specific project. Additionally, its value can be determined by identifying specific benefits that data modeling provides and then quantifying those benefits in all projects. Finally, data models can be reused in whole or in part for multiple projects, which can result in significant savings for any organization (Haughey 2010). Models also can be used to regenerate the missing data (Diamantopoulou et al. 2005).



² Environmental Engineering Department, KN Toosi University of Technology, P.O Box 1587-544-16, Tehran, Iran

It is better to make predictor models for the most important parameters, which in this case include:

- The dissolved oxygen (DO), which is an important quality index for some types of water. However, it is difficult to simulate the DO concentration by traditional mathematical methods due to the different factors which affect different kinds of water (Lihua et al. 2008);
- 2. Biological oxygen demand (BOD), which is a major parameter used to determine the degree of pollution in effluents (Akilandeswari and Adline 2013);
- 3. Nitrate and ammonium ions, which estimating their concentration in surface and ground water is extremely important. These values are usually determined at laboratories using sophisticated equipment, with a turnaround time varying from 2 h to 3 days. In many instances, the results are needed at the site as quickly as possible (Rich et al. 2006);
- Measuring alkalinity is an important factor in determining a stream's ability to neutralize acidic pollution resulting from rainfall or wastewater. It is also one of the best measures of the sensitivity of the stream to acidic inputs (US EPA 2013);
- 5. Hardness is also an important factor in aquaculture. Calcium and magnesium are the most common sources of water hardness. (Wurts 2002).

In line with the above facts, such parameters as DO, BOD, hardness and alkalinity are chosen as targets of our modeling study. For complete coverage, the additional parameter of chloride (Cl) concentration is added to the list above. Chloride ion is the predominant natural form of chlorine and is extremely soluble in water. Major sources of chloride in natural waters are sedimentary rocks that particularly evaporate. Igneous rocks contribute only a fraction of total chloride. Other sources are industrial and domestic wastewater (Pradhan and Pirasteh 2011). This study was conducted as a part of a master thesis research at KNTU, Environmental Eng. Dept. (2014).

Methods and materials

It was decided to make a simulation model with seven parameters. Relations between different physical properties of these parameters represent one of the most fascinating problems in modern science. Since this is a fundamental scientific problem, answering it will also provide answers to other practical needs, especially when one property is easier to measure than another (Sevostianov and Shrestha 2010). Now, it is needed to choose a number of related parameters as inputs for these models. It is better to choose parameters as inputs that are most economic and easy to assay. Temperature and pH are two of the easiest and the least costly (almost without any cost) parameters to measure. Electrical conductivity (EC) is also another parameter that can be measured quite easily and almost with no cost. It also has good relationship with other output parameters. Many researchers have used EC as input data, for example (Sevostianov and Shrestha 2010), and many others. The following table shows the names of a number of these researchers.

After finding the most appropriate input and output, it was decided to use artificial neural networks (ANN) to make necessary models. Applications of ANN are increasing because they are capable of solving and modeling every kind of complicated problems. Also, they are able to complete and cover all kinds of required parameter data series. Application of ANN is also increasing in resolving optimization problems (Koncsos 2010).

Neural networks

The artificial neural network (ANN), as its name implies, is a technique that simulates the functions of human brain during the problem-solving process. Just as humans use knowledge gained from experience to new problems or situations, the structure of a neural network can be used in powerful computation of complex nonlinear relationships (Kuo et al. 2004). MLF (multilayer feed-forward) networks trained with back-propagation algorithm are the most popular type of networks (Svozil et al. 1997 and Koncsos 2010). For example, the following table contains information on some modeling papers. These papers used feedforward networks to make their own models. You can also see a three-layer MLF network in Fig. 2.

To understand the structure of networks processing, the structure of a neuron (which is the smallest and basic element for any type of ANN) should be known at first. Figure 1 depicts a simple neuron.

In this schema, p is the matrix of input data multiplied by w, which is the matrix of weight. The calculated data (matrix) are summed with b (the threshold coefficient) which can be understood as a weight coefficient for the connection with formally added neuron where a = 1 (socalled bias) (Svozil et al. 1997). The result of the sum will be a matrix called n. The output of the neuron will be the



Fig. 1 Simple neuron



"*a*" matrix, expressed as (Koncsos 2010; Svozil et al. 1997):

$$a = f(n) = f(b + wp) \tag{1}$$

A network consists of a number of layers that each layer contains a number of neurons. There are three types of layers (Rounds 2002; Svozil et al. 1997; Menhaj 2008):

- 1. The input layer contains the input data and defines them for the network
- The hidden layer(s) where the process is carried out. The number of hidden layers and the number of neurons in each layer are variables;
- 3. The output layer that represents the network results (a_i) .

Figure 2 shows a four-layer feed-forward network. The number of hidden layer(s) and the number of neurons in each layer are chosen by designer of the network and depend on the conditions of the data and design details.

Training means to change the weights (w_i) and biases (b_i) in order to get closer to the answers. There are two main types of training processes: supervised and unsupervised training. In supervised training [e.g., multilayer feedforward (MLF) neural network], the neural network knows the desired output, and adjustment of weight coefficients is



Fig. 2 Feed-forward network with five layers

Fig. 3 Three-layer feed-

forward network

done in a way that the calculated and desired outputs are as close as possible. In unsupervised training, the desired output is not known (Svozil et al. 1997). This type of networks is mostly used for division problems. In the supervised mode, the square of the difference between the network output (a_i) and the data output is assumed as the main criterion for estimating the rate of learning (α). Mean-squared error (MSE) is calculated as: (Ghaffari et al. 2006; Menhaj 2008).

mse
$$=$$
 $\frac{1}{m} \sum_{i=1}^{m} e^2 = \frac{1}{m} \sum_{i=1}^{m} (t_i - a_i)^2$ (2)

In order of the training carried out correctly, the process should be repeated until the required precision is reached. In the following procedure, every time the process repeats, weights and biases will change. Figure 3 shows w_{ij} , b_{ij} and transfer function of each layer in a three-layer feed-forward network. The calibration process for w_i , b_i (s) is as (Svozil et al. 1997; Abraham 2005; Menhaj 2008):

$$w_{i,j}^{(l+1)} = w_{i,j}^{(l)} - \alpha \frac{\partial e(w,b)}{\partial w_{i,j}^{(l)}}$$
(3)

$$b_{i,j}^{(l+1)} = b_{i,j}^{(l)} - \alpha \frac{\partial e(w,b)}{\partial b_{i,j}^{(l)}}$$
(4)

$$\frac{\partial e(w,b)}{\partial w_{i,j}^{(l)}} = \left[\frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial w_{i,j}^{(l)}} e\left(w,b;x^{(i)},y^{(i)}\right)\right] + \alpha w_{i,j}^{(l)}$$
(5)

$$\frac{\partial e(w,b)}{\partial b_{i,j}^{(l)}} = \left[\frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial b_{i,j}^{(l)}} e\left(w,b;x^{(i)},y^{(i)}\right)\right] + \alpha b_{i,j}^{(l)} \tag{6}$$

As shown in Table 1, most of the works done in the field of modeling assumed a number of data as inputs. Some clues about choosing the input data (in relation to choosing the least expensive and the most accessible parameters as





References	Input(s)	Output(s)
Patki et al. (2013)	Alkalinity, hardness, TS and MPN	WQI
Rak (2013)	Psychochemical parameters of interim water	Turbidity
Chitsazan et al. (2013)	Rain data, mean monthly temperature, relative humidity, discharge of irrigation canal, groundwater recharge from the plain boundary	Groundwater depth
Nejadkoorki and Baroutian (2011)	Meteorological and gaseous pollutants from different air quality monitoring stations	Maximum PM10 concentration
Gustavo Andres Cuesta Cordoba Ing (2011)	Temperature, pH, flow, pipe material, diameter, and age of pipes	Free chlorine
Zhang et al. (2010)	Temperature, BOD, NH3-N, COD	DO
Panda Rabindra et al. (2010)	Water-level date and time	Water-level data
Anctil et al. (2009)	Number of 12 parameter such as Q, P, F	Daily nitrate-nitrogen and suspended sediment fluxes
Kim and Gilley (2008)	Runoff, electrical conductivity (EC) and pH	DP and NH4–N
Jalili Ghazi Zade and Noori (2008)	Weekly amount of solid waste	Generated solid waste
Nadiri (2007)	Groundwater level	Groundwater level
Diamantopoulou et al. (2005)	Temperature, flow, EC, HCO3, SO4, Na, Cl, Ca and DO	Nitrate
Rounds (2002)	Air temperature, solar radiation, rainfall and stream flow	DO

Table 1 Information on some modeling researches

Table 2 Units and limits of input data

W.Q. parameters	Unit	Limits			
Т	°C	<u>≤</u> 3			
pH	-	6.5-8.75			
EC	µS/cm @25 °C	50-900			

inputs) are already mentioned. Useful tips about the structure of the network's training algorithm are provided as following:

- The selected input parameter should be related to the output (including the result and the goal of modeling). Closer relationship between the input and output data can help us to minimize the error. However, the reverse is also true, which means if some parameters chosen as input that are not related to the output, it would cause disorder in the training process, which will not help in the modeling.
- 2. By using a higher number of related data as input, the training will be longer and will need more neurons as well. However, it can also make our model more accurate.

Table 2 shows the parameters that we have chosen as inputs for each model. The MATLAB program can be used to do calculations (Chitsazan et al. 2013; Chu et al. 2013). All needed formulas are incorporated in this program. At first, one must define the design parameters of the network. It should be noted that excessive use of hidden neurons will

cause over-fitting, which means that the neural networks may overestimate the complexity of the target problem. It also greatly degrades generalization capability of the model, which can lead to significant deviation in predictions. In this case, determining the proper number of hidden neurons to prevent over-fitting is critical in function approximation using neural networks. There are various approaches to building the network in a constructive or destructive way, but the most common methods to determine whether a certain number of hidden neurons are optimal, are cross-validation and early-stopping (Setiono 2001). Actually, to get the best result, the number of layers and neurons should be chosen in proportion to the complexity of problems and the number of input and output parameters (Abraham 2005).

Data resources

The River time series' data are used, which have been received from 210 monitoring stations from all over Ireland. These data are available on the website of the Ireland EPA (www.Water Quality Environmental Protection Agency, Ireland.htm under the name of Raw River Data for SE 2012).¹ The data that are chosen in our work came from 3001 groups of existing data set, while each data set contains 10 parameters. Fig. 4 shows data such as *t*, pH, EC, DO, BOD, Cl, nitrate, alkalinity and TDH that are used in the current study.



¹ www.water.epa.gov/type/rsl/monitoring/vms510.cfm.



Fig. 4 T, pH, EC, DO%, BOD, N, Cl, TDS and alkalinity data used in the study

Results and discussion

As it is mentioned before, the most available parameters are chosen as the input data, and they are shown in Table 2. To increase the precision of our models, the limits of each data should be observed. For example, for nitrate, the limits of existing data are between 0 and 0.17 mg/l, but actually just 70 items of data are between 0.04 and 0.17 mg/l; hence, they are removed to prevent their misleading effect on the training process; otherwise, they could reduce the precision of the model. Therefore, all of the existing data are applied as shown in Table 2.

In this study, seven models are developed, each of which estimates one of the water quality parameters. All models are trained by the feed-forward back-propagation algorithm. The learning function for all models is the Levenberg–Marquardt (LM), and the transfer function for all models is the tansign(x) function. The number of layers and neurons (in each layer) is shown in Table 3. To verify the current model(s), all data which were taken into account for training and developing the model(s) are used. Furthermore, in order to confirm the precision of each model, three criteria of MSE, MAE and R factors are evaluated. Therefore, the following results are obtained:

$$E_i = D_i - M_i \tag{7}$$

where Di = i'th real data, Mi = i'th estimated data

$$MSE = \frac{1}{n} \sum_{i=1}^{n} E_i^2$$
(8)

Table 3 Network properties	W.Q. parameters	Unit	Limits	Used data (n)	Layers	Neurons in layers	MSE	MAE	R
	DO	mg/l	7–15	2930	5	2-4-1-1	0.92	0.71	0.93
	DO percentage	%	70–130	2917	4	32-32-16	47.56	5.15	0.95
	BOD	mg/l	1–3	1455	3	32-25	0.22	0.37	0.77
	Cl	mg/l	5-40	2924	4	8-32-8	19.95	3.09	0.82
	Alkalinity	mg/l	20-350	2768	4	16-32-32	542.9	15.4	0.85
	Total hardness	mg/l	20–400	2819	4	32-32-16	363.3	11.16	0.92
49474543413937353331292725232119171513119 7 5 3 1 Number of data				13 12 11 10 9 4947454341393735333129272523211917151311 9 7 5 3 1 Number of data					
4947454341393735333129272523	211917151311 9 7	5 3 1	2.9 2.4 (1) 1.9 (1) 1.4 (1) 0.9	4947454341393	73533312	92725232119171513	11 9 7		33 30 27 24 21 8 8 15 22 12 mg/lit)
Number of d	ata 211917151311 9 7 5 ata	5 3 1	240 210 180 120 90 60 30	4947454341393	73533312	mber of data	1975	25 21 17 13 90 50 3 1	Total hardne ss(mg/lit)

Fig. 5 Package results with pilot data comparison; solid line represents data value, and dash line represents model prediction value

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |E_i| \tag{9}$$

$$R = \frac{1}{n} \sum_{i=1}^{n} 1 - \frac{E_i}{D_i}$$
(10)

The number of data used for modeling of water quality parameters and their limits is shown in Table 3. The only exception is that the developed model for DO concentration only uses T as the input data. As Table 3 shows, the precision of all these models is very high and these models are, therefore, very reliable. Moreover, inasmuch as a data set of 210 monitoring stations is used for development of the current models, these models can be adapted to various conditions and can be used under different conditions too. As it is noticed, three parameters such as T, pH and EC are used as input data because they can be easily measured. Therefore, these models are very practical and can be used in any field related to the surface water quality assessment. Each model is tested with 50 randomly selected data sets to show the precision of the package, and the results are shown in Fig. 5. Simulation results of the proposed ANN models (Fig. 5) show reliable and high correlation results for the proposed ANN models.

Conclusion

It is known that ANN model can be used in many practical and scientific subjects. In this work, the focus of attention is on the ANN model and its ability to simulate river water quality data. The results show useful applications of ANN modeling as: reliable replacement for salinity test; controlling equipment and operators; proper tool for estimating the missing data; calibration of the measurement tools; the ability of predicting quality data; ability of performing sensitivity analyzes on the generated data by model for scientific applications; and suitable for the conditions with experimental difficulties.

Acknowledgments The authors are grateful to Dr Sohrab Soori for their editorial and revision assistance. Also, they are thankful of "Water Quality Environmental Protection Agency, Ireland," for providing data sets.

References

- Abraham A (2005) Artificial neural networks. Oklahoma State University, Stillwater, pp 901–908
- Akilandeswari S, Adline MH (2013) Prediction of BOD values in engineering work industrial effluent by Anfis modeling. Int J Res Pure Appl Phys 3(2):7–9
- Anctila F, Filion M, Tournebizeb J (2009) A neural network experiment on the simulation of daily nitrate-nitrogen and

suspended sediment fluxes from a small agricultural catchment. Ecol Model 220:879–887

- Chitsazan M, Rahmani R, Neyamadpour A (2013) Groundwater level simulation using artificial neural network: a case study from Aghili plain, urban area of Gotvand, south-west Iran. JGeope 3(1):35–46
- Chu HB, Lu WX, Zhang L (2013) Application of artificial neural network in environmental water quality assessment. J Agric Sci Technol 15:343–356
- Diamantopoulou MJ, Antonopoulos VZ, Papamichail DM (2005) The use of a neural network technique for the prediction of water quality parameters of Axios River in Northern Greece. Eur Water 11(12):55–62
- Donohue I, Irvine K (2008) Quantifying variability within water samples: the need for adequate subsampling. Water Res 42:476–482
- Ghaffari A, Abdollahi H, Khoshayand MR, Bozchalooi IS, Dadgar A, Rafiee-Tehrani M (2006) Performance comparison of neural network training algorithms in modeling of bimodal drug delivery. Int J Pharm 327:126–138
- Gustavo Andres Cuesta Cordoba Ing (2011) Using of artificial neural network for evaluation and prediction of some drinking water quality parameters within a water distribution system. Water management and water structures, Juniorstav, pp 1–11
- Haughey I (2010) The return on investment (ROI) of data modeling. CA, Erwin, March, pp 1–18
- Jalili Ghazi Zade M, Noori R (2008) Prediction of municipal solid waste generation by use of artificial neural network: a case study of Mashhad. Environ Res 2(1):13–22
- Kim M, Gilley JE (2008) Artificial Neural Network estimation of soil erosion and nutrient concentrations in runoff from land application areas. Comput Electron Agric 64:268–275
- Koncsos T (2010) The application of neural networks for solving complex optimization problems in modeling. In: Conference of Junior Researchers in Civil Engineering pp 97–102
- Kuo YM, Liu CW, Lin KH (2004) Evaluation of the ability of an artificial neural network model to assess the variation of groundwater quality in an area of black foot disease in Taiwan. Water Res 38:148–158
- Lihua C, Shengquan M, Li LI (2008) A model to evaluate do of river based on artificial neural network and style book. J Hainan Normal Univ Nat Sci 21(4):372–376
- McKnighta S, Fundera SG, Rasmussenb JJ, Finkelc M, Binninga PJ, Bjerga PL (2010) An integrated model for assessing the risk of TCE groundwater contamination to human receptors and surface water ecosystems. Ecol Eng 36:1126–1137
- Menhaj MB (2008) Fundamental of neural network, vol 1. Industrial Amir Kabir University, Tehran
- Nadiri A (2007) Predicting groundwater level surrounding Tabriz city. Msd. Thesis, Tabriz University
- Nejadkoorki F, Baroutian S (2011) Forecasting extreme PM10 concentrations using artificial neural networks. J Environ Res 6(1):277–284
- Panda Rabindra K, Pramanik N, Bala B (2010) Simulation of river stage using artificial neural network and MIKE 11 hydrodynamic model. Comput Geosci 36:735–745
- Patki VK, Shirihari S, Manu B (2013) Water quality prediction in distribution system using Cascade feed forward neural network. Int J Adv Technol Civil Eng, ISSN: 2231–5721, 2(1):84–91
- Pradhan B, Pirasteh S (2011) Hydro-chemical analysis of the ground water of the basaltic catchments: upper bhatsai region, Maharashtra. Open Hydrol J 5:51–57
- Rak A (2013) Water turbidity modelling during water treatment processes using artificial neural networks. Int J Water Sci 2(3):1–10



- Rich D, Washo BD, Paladini A (2006) Rapid field test for nitrate and ammonia in reclaimed water. Everglades Res Educ Center 2:2006
- Rounds SA (2002) Development of a neural network model for dissolved oxygen in the Tualatin River. In: Oregon Second Federal Interagency hydrologic modeling conference, Las Vegas, Nevada, July 29–August 1, pp 1–13
- Schleiter IM, Borchardt D, Wagner R, Dapper T, Schmidt KD, Schmidt HH, Werner H (1999) Modeling water quality, bioindication and population dynamics in lotic ecosystems using neural networks. Ecol Model 120:271–286
- Scholten H, Kassahun A, Refsgaard JC, Kargas T, Gavardinas C, Beulens AJM (2007) A methodology to support multidisciplinary model-based water management. Environ Model Softw 22:743–759
- Setiono R (2001) Feed-forward neural network construction using cross validation. Neural Comput 13(12):2865–2877
- Sevostianov I, Shrestha M (2010) Cross-property connections between overall electric conductivity and fluid permeability of

a random porous media with conducting skeleton. Int J Eng Sci 48:1702–1708

- Stockholm International Water Institute and Elsevier (2012) The water and food nexus: trends and development of the research landscape
- Svozil D, KvasniEka V, Pospichal J (1997) Introduction to multilayer feed-forward neural networks. Chemometr Intell Lab Syst 39:43–62
- United States Environment Protection Agency (2013) Total Alkalinity. Retrieved 6 Mar 2013
- Varnell LM, Evans DA, Bilkovic DM, Olney JE (2008) Estuarine surface water allocation: a case study on the interactive role of science in support of management. Environ Sci Policy 11:602–612
- Wurts WA (2002) Alkalinity and hardness in production ponds. World Aquac 33:16–17
- Zhang Z, Wang X, Ou Y (2010) Water simulation method based on BPNN response and analytic geometry. Proc Environ Sci 2:446–453