An MLP-ANN-based approach for assessing nitrate contamination

Maria Laura Foddis, Augusto Montisci, Fatma Trabelsi and Gabriele Uras

ABSTRACT

This paper investigates the feasibility of predicting nitrate contamination from agricultural sources using multi-layer perceptron artificial neural networks (MLP-ANNs). The approach consists in training an MLP-ANN to predict nitrate concentrations based on a set of indirect measurements, such as pH, electrical conductivity, temperature and groundwater level. These are simpler and more economical than direct measurements, and they can be continuously collected on-site, rather than by performing laboratory tests. The approach has been validated in the nitrate vulnerable zone of the Arborea plain (central western Sardinia, Italy) by comparing the results obtained with different MLP-ANN models in order to find the most efficient model. The results show that the MLP-ANN-based model is a time- and cost-efficient method for predicting nitrate concentration.

Key words | artificial neural networks, multi-layer perceptron, nitrate contamination

Maria Laura Foddis (corresponding author)
Gabriele Uras
Department of Civil, Environmental Engineering and Architectural – Sector of Applied Geology and Applied Geophysics,
University of Cagliari,
via Marengo 3, 09123 Cagliari,
Italy
E-mail: *ing.foddis@gmail.com*

Augusto Montisci
Department of Electrical and Electronic Engineering,
University of Cagliari,
via Marengo 3, 09123 Cagliari,
Italy

Fatma Trabelsi
Higher School of Engineers of Medjez El Bab,
University of Jendouba,
Tunisia

INTRODUCTION

Groundwater is an important resource exploited for human consumption, and agricultural and industrial activities. One of the common kinds of pollution, in both surface and groundwater, that may affect agricultural and industrial areas, is nitrate contamination. To tackle this form of pollution, in 1991 the Council of the European Communities promulgated European Nitrates Directive 91/676/EEC with the objective of reducing water pollution caused or induced by nitrates from agricultural sources (fertilizers and organic slurry) and preventing contamination from further increasing by defining Nitrate Vulnerable Zones (NVZs). Nitrate contamination is a major issue in the region of Sardinia, particularly in the Arborea plain NVZ where intensive agriculture practices and farming are the major pillars of the local economy. Due to this leading role, agricultural practices have seen a substantial increase in the use of chemical and organic fertilizers to enhance productivity. In addition, the intensive use of groundwater for irrigation has caused the deterioration of groundwater and, in particular, the surficial aquifer, making the area more vulnerable to nitrate contamination. In fact, in the Arborea NVZ, groundwater nitrate concentration very often exceeds the water quality standard of 50 mg/l set by the Drinking Water Directive 80/778/EEC and its 98/83/EC revision.

A variety of studies have been conducted in the Arborea plain to determine the hydrological features of the area (Ghiglieri *et al.* 2016), to explore saltwater intrusion phenomena (Barrocu *et al.* 2004), and to investigate the intrinsic vulnerability of the shallow aquifer (Foddis *et al.* 2015b) and nitrate contamination (Foddis *et al.* 2015b, 2017). Due to the complex hydrological system in the area, the hydrogeological domain with all its peculiar features still needs to be further explored. Consequently, to address the need to model and predict nitrate contamination in agricultural areas, further research should be directed towards innovative approaches that make it possible to predict nitrate concentration in an efficient manner, in terms of both time and cost.

A number of recent studies have focused on the use of artificial neural networks (ANNs) to examine their suitability for modelling the uncertainty and complexity

doi: 10.2166/ws.2019.066

inherent in environmental processes (Mohamed & Hawas 2004; Singh & Datta 2007; Yesilnacar *et al.* 2007; Nourani *et al.* 2014, 2017; Secci *et al.* 2015; Foddis *et al.* 2015a, 2015b, 2017). Little has been published on predicting nitrate concentration in groundwater using artificial neural networks. Zaqoot *et al.* (2018) compared the results of implementing two ANN algorithms for predicting nitrate concentration based on a set of seven water quality parameters determined from measurements and laboratory analyses. Their results show that ANN models are able to predict nitrate contamination with acceptable accuracy. Ostad-Ali-Askari *et al.* (2016) used an ANN to estimate nitrate in groundwater based on 12 parameters of water quality analysis. Their results confirm that ANN models can be employed for investigating water quality parameters. Sathish Kumar *et al.* (2013) applied an ANN for estimating nitrate in groundwater using a series of input parameters found to have a strong correlation with, and a major influence on, nitrate contamination. Mousavi & Amiri (2012) adopted an adaptive neural-based fuzzy inference system for estimating nitrate concentration on the basis of 11 water quality variables measured in the laboratory. Their results showed that increasing the number of input variables improves the accuracy of nitrate estimates.

This paper discusses the development of a methodology that adopts the multi-layer perceptron artificial neural network (MLP-ANN)-based approach to predict the concentration of nitrate in aquifers using easily and economically quantifiable parameters such as pH, electrical conductivity, temperature and groundwater level. As opposed to other works, the methodology allowed us to predict nitrate concentration in water wells using ground-measured data. Furthermore, the possibility of performing simple ground measurements enabled continuous remote monitoring of the aquifers. The procedure proposed has been validated using the measurements from a set of wells located in the Arborea plain NVZ.

METHODS

This section includes a general description of the MLP-ANNs and the method adopted to predict agricultural nitrate concentrations.

ANN design

In this paper, an MLP-ANN (Principe *et al.* 2000) is used to associate nitrate concentrations with the ground-measured data. As is well known, the most important feature of ANNs is their generalizability, namely their ability to properly approximate examples not included in the training set. As a general rule, with a large number of degrees of freedom the ANN works well with the training set, but at the expense of generalization ability. In contrast, if the number of examples is too small then they cannot be properly approximated. Several rules-of-thumb are proposed in the literature to determine the appropriate number of degrees of freedom, mainly based on the number of training examples. Nonetheless, this heuristic yields a very broad range of possible values from which to choose, as the optimal value depends on the actual distribution of the points, which is generally unknown. It has been demonstrated (Cybenko 1989) that an MLP with only one hidden layer is a universal approximator. On the other hand, in many cases it has been observed that two hidden layers of neurons improve MLP performance even with the same number of degrees of freedom. In this work, the number of neurons in the hidden layer has been assigned by means of incremental evolution (see 'Incremental evolution for hidden layer sizing' below).

Training strategy

The performance of a trained MLP critically depends on the training set. In fact, the training set needs to be representative of the entire distribution of possible cases. However, at the same time, the number of training examples should be maintained to be as small as possible, as this affects the number of degrees of freedom and ultimately computational cost. The main contribution of this paper concerns the procedure adopted for selecting the training set.

Another issue concerning training is overfitting. In practice, if the network is overtrained its generalization ability is reduced, and in spite of providing good accuracy on the training examples, with unacceptable errors on any other examples. To avoid this, a test set is defined, which is uncorrelated with the training set, and the approximation error is evaluated on both the training and test sets. If the latter exceeds the acceptance threshold, then training has failed.

The most common way to use the test set is early stopping (Principe *et al.* 2000), which consists in evaluating the error on the test set during the training phase, and in stopping when the error on the test set begins to increase. This is simply heuristics, and there is no guarantee that after a certain number of epochs, the test error will not begin to diminish again.

In light of the above considerations, in the present work the following strategy was implemented. First of all, the training set was selected from among all the available examples by prioritizing the *vertex* points, namely the points that, in the product space {Input}×{Output}, cannot be obtained as a convex combination of the remainder of the set (see 'Training set selection' below). The next step consisted in determining the size of the hidden layer and the number of layers. To this end, we adopted a growing procedure, training several ANNs with a growing number of hidden neurons, and for each case, the Mean Squared Error (MSE) was determined on both the training and test sets. Optimal size was taken as the best approximation of the test set (see 'Incremental evolution for hidden layer sizing' below). In order to deal with local minima problems, an appropriate number of ANNs, all having the same number of hidden neurons, but with different initial sets of weights, are trained in parallel. In calculating the output of a given example, the outputs obtained by the single ANN are combined by applying a weighted *majority voting* criterion (Bauer & Kohavi 1999) (see 'Majority voting to increase generalization ability' below). In the following three subsections, the procedure is described in detail.

Training set selection

In order to select the training set from among the available examples, the distribution of the whole set is examined on the product space {R} = {Input}×{Output}, namely the geometrical space that encompasses the input and output spaces. It is reasonable to assume that it is easier to interpolate a point inside the distribution than extrapolate outside it. Therefore it is worth assessing the performance of the system if the training set is composed of points that cannot be obtained as a convex combination of the rest of the points. In other words, all the convex combinations among points of the set form a polyhedron and we take as the training set the vertices of this polyhedron. Linear programming (Bazaraa *et al.* 2011) is used to establish whether a point is a convex combination of the current training set. This test is briefly described below. A point $\underline{P}$ is a convex combination of a set of points $\underline{Q}_k$, $k = 1, \ldots, K$ if the following equations hold:

$$\begin{aligned} \underline{P} &= \sum_k \alpha_k \underline{Q}_k \\ \sum_k \alpha_k &= 1 \\ \alpha_k &\geq 0 \qquad \forall k \end{aligned} \tag{1}$$

The points with an extreme value (maximum or minimum) of one coordinate are indeed vertices, and as such can be included without performing any tests. Starting from this initial set, test (1) is performed iteratively on the whole residual set, and at each iteration, only the point with the maximum value of the objective function is included in the training set.

Incremental evolution for hidden layer sizing

Although Cybenko (1989) demonstrated that an MLP with only one sigmoidal hidden layer represents a universal approximator, establishing the appropriate number of neurons in this layer is still an open issue. What we do know is that a large hidden layer facilitates learning but at the same time the performance of the test set deteriorates. Therefore, the best solution is determined by trial and error. Furthermore, even if a single hidden layer is sufficient to achieve an arbitrary degree of approximation, experience suggests that distributing the hidden neurons into two layers allows one to obtain better performance with the same total number of neurons. In this work, a number of MLPs each having one hidden layer of a different size have been trained, and their performance compared. Performance is evaluated on both the training and the test sets, and in terms of both MSE and the greatest error in the set.

Majority voting to increase generalization ability

A key issue in developing an MLP-ANN is its generalization capability, namely the ability to maintain a suitable degree

of approximation for those examples not belonging to the training set. To this end, a stopping criterion is adopted, based on a test set of examples independent of the training set. When the error in the test set begins to increase, training is interrupted (early stopping) (Principe *et al.* 2000). To ensure this approach is effective, it is important that both the training and test sets are representative of the whole population. As the performance of the ANN depends on the initial set of connection weights, a certain number of ANNs with the same size but with different initial weights are trained in parallel. Their outputs are combined by a weighted sum, where the weights are assigned on the basis of the reliability of the ANN. More specifically, the weight of each output is proportional to the inverse of the MSE of the ANN on the training set. The weights are normalized, so that their sum is equal to 1. This makes the system robust with respect to any outlier error made by an ANN on a single sample.

CASE STUDY

The proposed method has been evaluated by predicting nitrate concentration in 42 water wells throughout the NVZ of Arborea plain.

Study area

The study area is located in the Arborea plain (central western Sardinia, Italy) (Figure 1). The area extends over

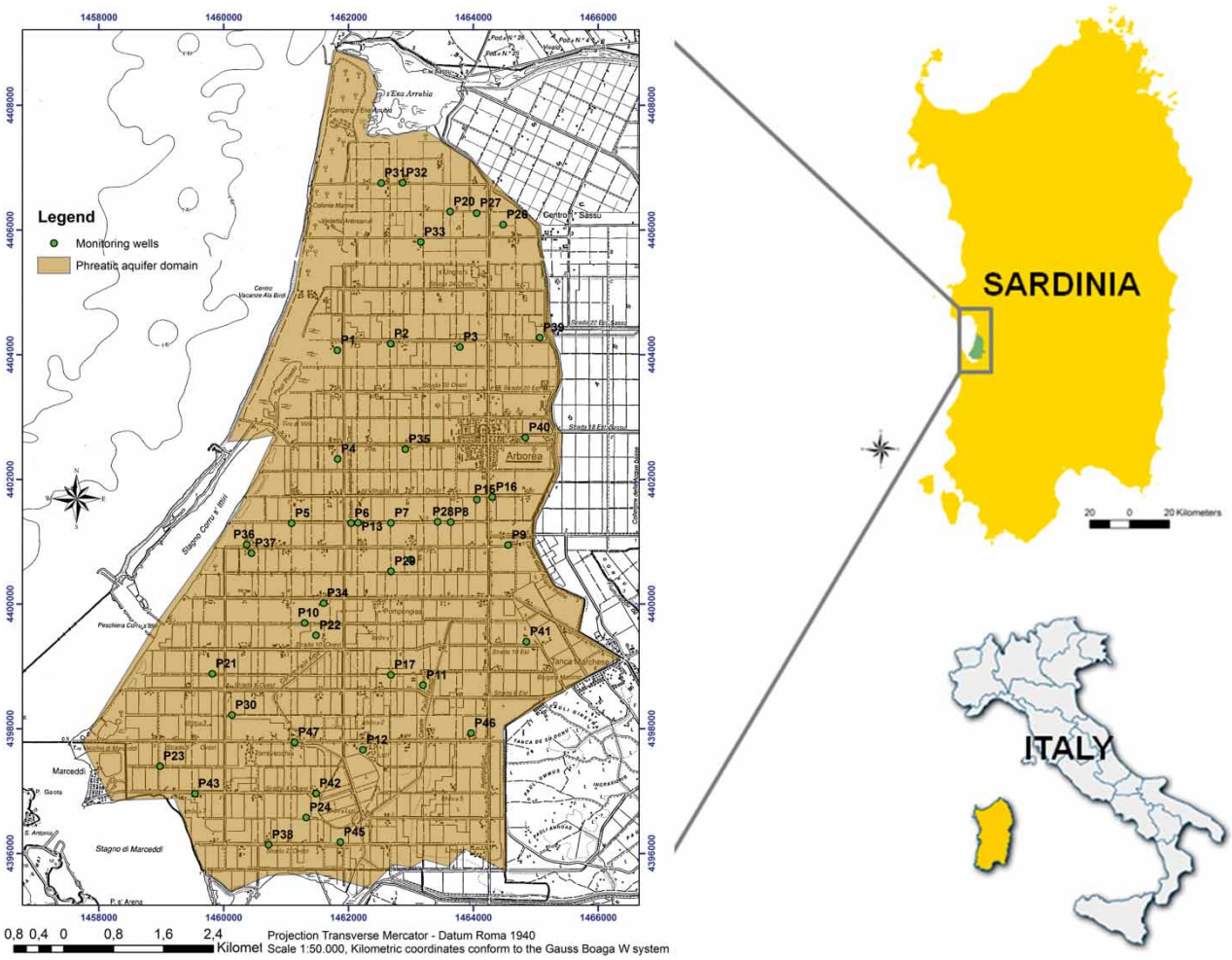

Figure 1 | Study area and water well location (NVZ of Arborea, Sardinia, Italy).

roughly 70 km^2 and forms part of the coastal flood plain overlooking the Gulf of Oristano. The area lies within the administrative boundaries of the municipality of Arborea and in the northernmost part of the Campidano plain.

The Arborea plain is one of the most productive agricultural areas in Sardinia, characterized by irrigated crops and livestock holdings. Intensive agriculture and dairy farming have resulted in this area carrying the risk of nitrate contamination. Agricultural practices include the use of chemical fertilizers and animal manure to increase yields. These practices, together with aquifer overexploitation, leaching and the infiltration of huge quantities of nitrogen into the ground, have made this area particularly vulnerable to nitrate pollution. Consequently, it has been designated NVZ in the Sardinian Water Safety Plan (Piano di Tutela delle Acque 2006) drawn up pursuant to EEC Nitrates Directive 91/676/EEC.

Data pre-processing

The 42 water wells located in the study area, monitored on a monthly basis by ARPAS (Sardinian Agency for Environmental Protection of Sardinia) (Figure 1), only intercept the shallow aquifer, water being withdrawn by local farmers almost daily. This investigation takes into account 482 measurements of nitrate concentration (NO_3), pH, electrical conductivity (EC), temperature (T), and groundwater level (GL) collected during the monitoring campaigns in different seasons from 2007 to 2011. Therefore the whole data set was composed of 482×5 values. In addition, the availability of samples acquired in different seasons and at geographical points is fundamental for guaranteeing the representativeness of the data set used to train the ANN model. During a preliminary phase the linear correlation, correlation coefficient and the significance of data correlation were evaluated for the data coupled in input–output pairs (GL–NO_3; T–NO_3; EC–NO_3; pH–NO_3) and input–input pairs (GL–NO_3; T–NO_3; EC–NO_3; pH–NO_3). This first analysis highlighted the lack of correlation between both pairs of variables. This first result is very interesting because non-linear correlation can provide more significant information for training purposes, avoiding signal redundancy, and at the same time it justifies the interest in employing a non-linear modelling tool such as MLP-ANNs.

Developing the MLP-ANN model

The measurement campaigns described in the subsection 'Study area' provided a total of 482 input–output pairs, where the input is represented by the measurements of pH, electrical conductivity, temperature, and groundwater level, while the output represents the nitrate concentration measurements performed in the laboratory. The structure of the MLP-ANN will be 4-*x*-1, which indicates that the MLP-ANN has one hidden layer, and the number of neurons has to be established.

First, the training set was determined, adopting the procedure described in the subsection 'Training set selection', obtaining a set of 65 examples. The remaining 417 examples formed the test set. The size *x* of the hidden layer was varied between 7 and 45. For each size, 20 MLPs with random initial weights were trained, and their performance combined by means of the majority voting method (majority voting to increase generalization ability). All the MLPs were trained for a fixed number of 60 epochs. For each size of the hidden layer, the performance of the corresponding group of 20 ANNs was evaluated by considering both the MSE and the maximum error on both the training and test sets. In the following section, the results are reported and commented.

RESULTS AND DISCUSSION

Figure 2 shows the effect of increasing the size of the hidden layer on the approximation level. As can be observed,

Figure 2 | Trend of errors vs number of hidden neurons.

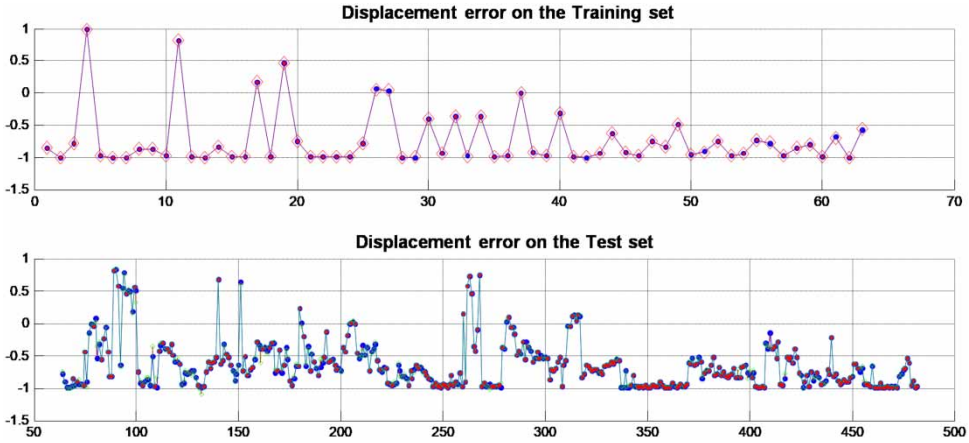

Figure 3 | Displacement of the output on the training and test sets.

increasing the number of hidden neurons does not affect the generalization ability of the MLP-ANN. In fact, the error on the test set, both the maximum error and the MSE, exhibits the same trend as the training set. This behaviour indicates the training set has been appropriately selected, as the small number of training examples is able to well represent the whole set, and overfitting is avoided. Note that the trends are not regular, in the sense that in certain portions the curve first rises, then declines. This behaviour cannot be attributed to a specific size of the MLP-ANN, but can rather be considered as fluctuations due to the initial randomization of the weights. By training several networks combined through majority voting the curve becomes smoother. Clearly, the MLP-ANN becomes increasingly accurate as size increases.

Figure 3 shows the resulting degree of approximation of the MLP-ANN 4-40-1, the largest one. As can be observed, the MLP-ANN provides good accuracy for both the training and test sets, demonstrating the suitability of the proposed approach for predicting nitrate concentrations.

CONCLUSION

This work focuses on an MLP-ANN-based methodology for estimating nitrate contamination from agricultural sources, based on a set of ground measurements. A key aspect of the procedure is MLP-ANN training, which needs to ensure a satisfactory approximation of the physical system as well as generalization ability. Nitrate contamination in water wells can be predicted on the basis of easily and economically quantifiable parameters employing an MLP-ANN model. Furthermore, the MLP-ANN approach has proven to be both time- and cost-efficient. The results show that the development of an MLP-ANN needs to be supported by algorithms designed and tested specifically for each case study, so as to be able to exploit a significant amount of information from the available data. MLP-ANNs may offer a valuable contribution to the pool of existing solutions for the control and abatement of nitrate contamination, by allowing the accurate monitoring of the progressive degradation of groundwater resources in NVZs and then the identification of action plans aimed at informing and training farmers to improve fertilization management and agricultural practices. To the best of our knowledge, no studies reported in the literature have adopted this approach for the remote monitoring of nitrate contamination in groundwater.

ACKNOWLEDGEMENTS

This article was based on the data collected by ARPAS, the Sardinian Agency for Environmental Protection.

REFERENCES

Barrocu, G., Cau, P., Soddu, S. & Uras, G. 2004 Predicting groundwater salinity changes in the coastal aquifer of Arborea (central–western Sardinia). In: *18th Salt Water*

Intrusion Meeting (SWIM) (L. Araguás, E. Custodio & M. Manzano, eds), Instituto Geologico y Minero de España, Madrid, Spain, pp. 243–255.

Bauer, E. & Kohavi, R. 1999 An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning* **36**, 105–139.

Bazaraa, M. S., Jarvis, J. J. & Sherali, H. D. 2011 *Linear Programming and Network Flows*. John Wiley & Sons, Hoboken, NJ, USA.

Cybenko, G. 1989 Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* **2**, 303–314.

Foddis, M. L., Ackerer, P., Montisci, A. & Uras, G. 2015a ANN-based approach for the estimation of aquifer pollutant source behaviour. *Water Science and Technology: Water Supply* **15** (6), 1285–1294.

Foddis, M. L., Matzeu, A., Montisci, A. & Uras, G. 2015b Application of three different methods to evaluate the nitrate pollution of groundwater in the Arborea plain (Sardinia – Italy). *Rendiconti Online Società Geologica Italiana* **35**, 136–139.

Foddis, M. L., Matzeu, A., Montisci, A. & Uras, G. 2017 The Arborea plain (Sardinia-Italy) nitrate pollution evaluation. *Italian Journal of Engineering Geology and Environment* **2017** (Special Issue 1), 67–76.

Ghiglieri, G., Carletti, A., Da Pelo, S., Cocco, F., Funedda, A., Loi, A., Manta, F. & Pittalis, D. 2016 Three-dimensional hydrogeological reconstruction based on geological depositional model: a case study from the coastal plain of Arborea (Sardinia, Italy). *Engineering Geology* **207**, 103–114.

Mohamed, A. M. O. & Hawas, Y. 2004 Neuro-fuzzy logic model for evaluating water content of sandy soils. *Computer-Aided Civil and Infrastructure Engineering* **19** (4), 274–287.

Mousavi, S. F. & Amiri, M. J. 2012 Modelling nitrate concentration of groundwater using adaptive neural-based fuzzy inference system. *Soil and Water Research* **7** (2), 73–83.

Nourani, V., Hosseini Baghanam, A., Adamowski, J. & Kisi, O. 2014 Applications of hybrid wavelet–artificial intelligence models in hydrology. *Journal of Hydrology* **514**, 358–377.

Nourani, V., Mousavi, S., Dabrowska, D. & Sadikoglu, F. 2017 Conjunction of radial basis function interpolator and artificial intelligence models for time-space modeling of contaminant transport in porous media. *Journal of Hydrology* **548**, 569–587.

Ostad-Ali-Askari, K., Shayannejad, M. & Ghorbanizadeh-Kharazi, H. 2016 Artificial neural network for modeling nitrate pollution of groundwater in marginal area of Zayandeh-rood River, Isfahan, Iran. *KSCE Journal of Civil Engineering* **21** (1), 134–140.

Piano di Tutela delle Acque (Water Safety Plan) approved by the Autonomous Region of Sardinia, resolution number 14/16 of 04/04/2006.

Principe, J. C., Euliano, N. R. & Lefebvre, W. C. 2000 Innovating adaptive and neural systems instruction with interactive electronic books. *Proceedings of the IEEE* **88** (1), 81–95.

Sathish Kumar, S., Mageshkumar, P., Santhanam, H., Stalin John, M. R. & Amal Raj, S. 2013 A new logic-based model to predict nitrates in groundwater using artificial neural network (ANN). *Pollution Research* **32** (3), 635–641.

Secci, R., Foddis, M. L., Mazzella, A., Montisci, A. & Uras, G. 2015 Artificial Neural Networks and Kriging method for slope geomechanical characterization. In: *Engineering Geology for Society and Territory – Volume 2* (G. Lollino, D. Giordan, G. B. Crosta, J. Corominas, R. Azzam, J. Wasowski & N. Sciarra, eds), Springer, Heidelberg, Germany, pp. 1357–1361.

Singh, R. M. & Datta, B. 2007 Artificial neural network modeling for identification of unknown pollution sources in groundwater with partially missing concentration observation data. *Water Resources Management* **21**, 557–572.

Yesilnacar, M. I., Sahinkaya, E., Naz, M. & Ozkaya, B. 2007 Neural network prediction of nitrate in groundwater of Harran Plain, Turkey. *Environmental Geology* **56** (1), 19–25.

Zaqoot, H. A., Hamada, M. & Miqdad, S. 2018 A comparative study of Ann for predicting nitrate concentration in groundwater wells in the southern area of Gaza Strip. *Applied Artificial Intelligence* **32** (7–8), 727–744.

First received 11 November 2018; accepted in revised form 11 April 2019. Available online 22 April 2019