Prediction of the quality of public water supply using artificial neural networks

Henrique Vicente, Susana Dias, Ana Fernandes, António Abelha, José Machado and José Neves

ABSTRACT

The Health Surveillance Program was established by the Regional Health Authority of Alentejo to control the quality of public water supply. This authority divides the water quality parameters into three distinct groups, namely P_1 (pH and conductivity), P_2 (nitrate and manganese) and P_3 (sodium and potassium), for which the sampling frequency is dissimilar. Thus, the development of formal models is essential to predict the chemical parameters included in group P_2 and included in group P_3, for which the sampling frequency is lower, based on the chemical parameters included in group P_1. In the present work, artificial neural networks (ANNs) were used to predict the concentration of nitrate, manganese, sodium and potassium from pH and conductivity. Different network structures have been elaborated and evaluated using the mean absolute deviation and the mean squared error. The ANN selected to predict the concentration of nitrate, sodium and potassium from pH and conductivity has a 2-18-14-3 topology while the network selected to predict the concentration of nitrate and manganese has a 2-19-10-2 topology. A good match between the observed and predicted values was observed with the R^2 values varying in the range 0.9960–0.9989 for the training set and 0.9993–0.9952 for the test set.

Key words | artificial neural networks, monitoring of public water supply, prediction of water quality parameters

Henrique Vicente (corresponding author)
Escola de Ciências e Tecnologia,
Departamento de Química e Centro de Química de Évora, Universidade de Évora,
Rua Romão Ramalho, 59, 7000-671 Évora,
Portugal
E-mail: *hvicente@uevora.pt*

Susana Dias
Administração Regional de Saúde do Alentejo IP,
Laboratório de Saúde Pública de Évora,
Hospital do Patrocínio – 4° Piso,
Av. Infante D. Henrique, 7000-811 Évora,
Portugal

Ana Fernandes
Escola de Ciências e Tecnologia,
Departamento de Química,
Universidade de Évora,
Rua Romão Ramalho, 59, 7000-671 Évora,
Portugal

António Abelha
José Machado
José Neves
Departamento de Informática,
Universidade do Minho,
Braga,
Portugal

INTRODUCTION

Ensuring the quality of water intended for human consumption is a major goal in contemporary societies, taking into account the importance to health and the need to safeguard and promote its sustainable use. In Portugal, Decree-Law N° 306/2007 of 27 August sets the quality of drinking water, defining 33 mandatory and 28 indicators parameters. It aims to protect human health from the adverse effects of possible contamination of the water and to ensure universal availability of clean water with a balanced composition.

Concerning the sanitary surveillance, Decree-Law N° 306/2007 of 27 August states that the Regional Health Authorities promote the regular and periodic sanitary surveillance of the water for public supply, establishing, for this purpose, the regional Sanitary Surveillance Program (SSP). The Public Health Laboratories (PHL) are responsible for the analytical support of the activities included in the SSP. The PHL of Évora, a district in the south of Portugal where this work took place, depends on the Regional Health Administration of the Alentejo and is accredited by the Portuguese Institute for Accreditation (IPAC), in accordance with standard NP EN ISO/IEC 17025 for tests in the field of chemistry and microbiology of water since 2006. Currently, the PHL of Évora carries out surveillance of the drinking water for the 14 municipalities of the District of Évora.

The SSP establishes the parameters to be analysed and the sampling frequency. Table 1 shows the parameters included in the SSP, divided into three groups, namely P_1,

doi: 10.2166/aqua.2012.014

Table 1 | Parameters included in the Sanitary Surveillance Program

	Group P_1	Group P_2	Group P_3
Microbiological parameters	Coliform bacteria *Escherichia coli* *Enterococci*	P_1 Microbiological parameters + *Clostridium perfringens*, Cultivable microorganisms at 22 and 37 °C	P_2 Microbiological parameters
Chemical parameters	pH, Conductivity	P_1 Chemical parameters + turbidity, ammonium, nitrate, nitrite, iron, manganese, oxidability	P_2 Chemical parameters + sodium, potassium, cadmium, chromium, lead, copper, nickel

P_2 and P_3. The sampling frequency of the former groups is depicted in Figure 1. As is shown in Table 1, most of the chemical parameters used in the SSP are included in groups P_2 and P_3, for which the sampling is less frequent (Figure 1). Thus, in order to control the quality of public water supplies in real time, the development of models to predict the concentrations of these parameters based on the chemical ones included in group P_1 (i.e. pH and conductivity) can become essential.

In recent years, some artificial intelligence based tools, namely artificial neural networks (ANNs) and decision trees, have been applied for water quality assessment (Santos *et al.* 2005; Pinto *et al.* 2009; Farmaki *et al.* 2010; Maier *et al.* 2010; West & Dellana 2011). However, the prediction of the concentration of nitrate, sodium, potassium and manganese using pH and conductivity is a complex and highly nonlinear problem.

The aim of the current study was to use ANNs to solve this problem. ANNs can learn from examples, are fault tolerant in the sense that they are able to handle noisy and incomplete data, are able to deal with nonlinear problems and, once trained, can perform prediction and generalization (Haykin 2008).

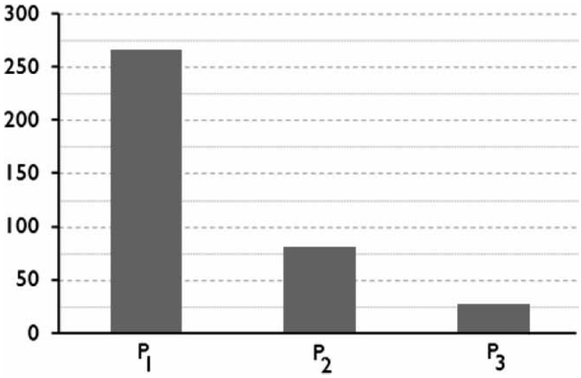

Figure 1 | Annual average sampling frequency.

MATERIALS AND METHODS

The water samples used for the development of the models were collected in the public water supply networks of the district of Évora during the time period from January 2007 to December 2009. The parameters analysed were pH, conductivity, oxidability, turbidity, ammonium, iron, nitrate, nitrite, manganese, cadmium, chromium, lead, copper, nickel, sodium and potassium.

Sample collection and preservation

Sample collection was carried out on the consumer's tap, following the procedures recommended by the Portuguese Water and Waste Management Service (IRAR 2005). Thus, without prior discharge, the first litre of water was collected in a polyethylene bottle, rinsed with nitric acid, for the analysis of lead, copper and nickel. Thereafter, samples were collected for the analysis of the remaining chemical parameters (IRAR 2005).

Sample preservation followed the procedures described in Standard Methods for the Examination of Water and Wastewater (SMEWW) (Eaton *et al.* 2005). For pH and conductivity, the samples were collected in 50 mL wide-mouth polyethylene bottles and analysed immediately; for oxidability and turbidity analysis, the samples were collected in polyethylene bottles of 100 mL, stored in the dark and kept refrigerated; for nitrate and nitrite analysis the samples were collected in polyethylene bottles of 100 mL and kept refrigerated; for ammonium analysis, the samples were collected in 500 mL polyethylene bottles, preserved with sulphuric acid, pH ≤ 2 and kept refrigerated; for iron analysis the samples were collected in a 100 mL glass bottle, preserved with sulphuric acid 4.5 M, pH ≤ 2; finally, for remaining metals analysis the samples were collected in 1,000 mL polyethylene bottles rinsed with nitric acid and preserved with nitric acid, pH ≤ 2.

Analytical procedures

The analyses of water quality parameters followed the SMEWW (Eaton *et al.* 2005), the International Standard Organization (ISO), the European Standards or the Portuguese Standards (Table 2).

The evaluation of pH was accomplished using a Sherwood SCI Delta 345 pH meter equipped with a Mettler Toledo Inlab 412 electrode. The conductivity measurements were carried out on a Crison 2202 micro CM conductivity meter equipped with a Crison ACC 5292 cell. The turbidity assessments were carried out on an HF Scientific Micro turbidimeter. The molecular absorption spectrometry evaluation was accomplished on a Thermo Electron spectrometer model Nicolet Evolution 300 LC. The atomic absorption spectrometry measurements were performed on a Perkin Elmer 3110 spectrometer equipped with a HGA-600 graphite furnace. Finally, the flame photometry measurements were completed on a Corning 410 photometer.

Artificial neural networks

ANNs are computational tools inspired by the architecture and the internal operational features of the human brain (Haykin 2008). In this study the most common neural network type, the multilayer perceptron, was adopted. This type of network is formed by three or more layers of basic computing units named artificial neurons or nodes. It includes an input layer, an output layer and a number of hidden layers with a certain number of active neurons connected by feedforward links, to which are associated modifiable weights. In addition, there are also bias, which are connected to neurons in the hidden and output layers. The number of nodes in the input layer denotes the number of independent variables and the number of nodes

Table 2 | Analytical techniques, test methods and detection limits

Parameter	Analytical technique	Test method	Detection limit
pH	Potentiometry	SMEWW 4500-H^{+}	–
Conductivity	Conductimetry	NP EN 27888:1996[a]	26.0 μS/cm
Oxidability	Volumetry	NP 731:196[b]	0.9 mg/dm^{3}
Turbidity	Turbidimetry	ISO 7027:1999	0.25 UNT
Ammonium	Molecular absorption spectrometry	ISO 7150–1:1984	0.15 mg/dm^{3}
Iron		NP 2202:1996[b]	0.05 mg/dm^{3}
Nitrate		SMEWW 4500-NO$_3^-$	1.5 mg/dm^{3}
Nitrite		NP EN 26777:1996[a]	0.012 mg/dm^{3}
Manganese	Atomic absorption spectrometry	SMEWW 3113-B	5.0 μg/dm^{3}
Cadmium			0.50 μg/dm^{3}
Chromium			5.0 μg/dm^{3}
Lead			6.0 μg/dm^{3}
Copper			6.0 μg/dm^{3}
Nickel			5.0 μg/dm^{3}
Sodium	Flame photometry	SMEWW 3500-Na	1.0 mg/dm^{3}
Potassium		SMEWW 3500-K	0.10 mg/dm^{3}

[a] NP EN – Portuguese version of European Standard.

[b] NP – Portuguese Standard.

in the output layer stands for the number of dependent variables (Haykin 2008).

Although it has been proven that a network with one hidden layer can approximate any continuous function, given sufficient degrees of freedom (Hornik *et al.* 1989), other studies have shown that, in practice, many functions are difficult to approximate with one hidden layer (Cheng & Titterington 1994; Flood & Kartam 1994). Indeed, there are no clear rules as to the 'best' number of hidden ones. Network design is a trial-and-error process and may affect the accuracy of the resulting trained network. A number of automated techniques have been proposed to search for a 'good' network structure.

These typically use a hill-climbing approach that starts with an initial structure that is selectively modified to improve performance, i.e. in order to minimize an internal measure of error. The internal error metric most commonly used is the mean squared error (MSE) (Cortez *et al.* 2004; Han & Kamber 2006).

In the training phase, the back propagation (BP) algorithm (Rumelhart *et al.* 1986) was applied. This is the most widely used training algorithm for the multilayered perceptron, and evolves in two phases. The former is concerned with the forward stage, where the information is propagated from the input to the output layer. The last is the backward stage where an error, defined as the discrepancy between the observed value and the desired nominal value in the output layer, is propagated backwards in order to adjust the weightings and bias values. In the forward phase, the weighted sum of input components, u_j, is calculated as:

$$u_j = \sum_{i=1}^{n} w_{ij} x_i + \text{bias}_j \qquad (1)$$

where w_{ij} denotes the weight between the jth and the ith neurons in the preceding layer, x_i stands for the output of the ith neuron in the preceding layer, and bias_j alludes to the weight between the jth neuron and the bias neuron in the preceding layer.

The output of the jth neuron in any layer, y_j, is given in the form:

$$y_j = f(u_j) \qquad (2)$$

where f refers to the activation function. In all experiments the sigmoid activation function was used, excluding the nodes in output layers. In this case the linear activation function was used.

The BP algorithm is controlled by two parameters, the momentum coefficient and the learning rate, ranging between 0 and 1. The momentum coefficient is used at the weights updating stage and tends to keep the weight changes evolving in a consistent direction. Learning rate controls how much the weights are adjusted at each update. The software used to implement ANNs was the Waikato Environment for Knowledge Analysis (WEKA), keeping the default software parameters (Hall *et al.* 2009).

To ensure statistical significance of the attained results, 20 runs were applied in all tests. In each simulation, the available data were randomly divided into two mutually exclusive partitions: the training set, with two-thirds of the available data, used during the modelling phase, and the test set, with the remaining one-third of the examples used after training in order to evaluate the model performance (Souza *et al.* 2002). To improve the performance of the learning algorithm and avoid the overvaluation of the attributes with larger intervals at the expense of the attributes with smaller ones, the data were normalized to the interval [0,1], using the equation depicted below (Han & Kamber 2006):

$$\overline{X} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \qquad (3)$$

where $\overline{X}$ alludes to the normalized value, X denotes an attribute value, and $X_{\min}$ and $X_{\max}$ stand for, respectively, the minimum and the maximum values that may be assigned to an attribute.

RESULTS AND DISCUSSION

Database

The data were obtained by the Public Health Laboratory of Évora and covers the time period from January 2007 to December 2009. However, once the sampling was carried out on the consumer's tap, the parameters turbidity, ammonium, nitrite, iron, cadmium, lead, copper, chromium and nickel always exhibited values below the detection limit of the analytical methods used in their quantification being, therefore, excluded. The database used in this study contained a total of 112 records with six fields. The fields were pH, conductivity and the concentrations of nitrate, sodium, potassium and manganese. Unfortunately, the number of complete records was small. The values for pH, conductivity and nitrate are always present but, conversely, if the values for manganese are present, there are not values for sodium and potassium and vice versa. Thus, two different databases were constructed. One of them was used to build predictive models for nitrate, sodium and potassium concentrations (Database A). The other was used to build predictive models for nitrate and manganese concentrations (Database B). Table 3 shows the statistical characterization of the fields included in the databases. Figure 2 depicts the relationships between dependent (nitrate, potassium, sodium and manganese) and independent variables (pH and conductivity).

Excluding pH, Table 3 shows large dispersion of the data with a high coefficient of variation, ranging from 32 to 85%. Such variability may be attributed to two main factors. One of them is the large geographical area of the district of Évora (7,393 km^2). The other has to do with the multiple water sources used, which include surface water and/or groundwater. There are three different hydrographic basins (i.e. Tagus River, Guadiana River and Sado River) where there are several reservoirs used for drinking water production. Concerning groundwater there are two main aquifer systems. The aquifer system Estremoz–Cano, with 202 km^2, in which calcium–magnesium–bicarbonate is the dominant hydrochemical facies, and the aquifer system Viana do Alentejo–Alvito, with 18.4 km^2, where the dominant hydrochemical facies are calcium–bicarbonate, magnesium–bicarbonate and calcium–magnesium–chloride (Almeida *et al.* 2000). The pH shows the coefficient of lowest variation, and it may be due to the pH adjustment that occurs in the water treatment plants. Nevertheless, these results are in agreement with results presented by other authors for similar systems (Palani *et al.* 2008; Singh *et al.* 2009).

Artificial neural network models

In order to obtain the best prediction of the output parameters (i.e. nitrate, sodium and potassium, on the one hand, and nitrate and manganese, on the other hand), different network structures and architectures were elaborated and evaluated. The optimum number of hidden layers and the optimum number of nodes in each of these cases was found by trial and error. Common tools to compare the performances of regression models are the mean absolute deviation (MAD) and the MSE. According to Torgo (1999), when these tools are applied to the evaluation of regression models serve different purposes. If the goal is a model that avoids large deviations, MSE should be minimized since this error metric amplifies large deviations. Conversely, if some large deviations can be allowed, the amplification effect is not necessary and MAD should be used. These two measures of goodness-of-fit are related to the average prediction error. Nevertheless, they do not provide any information on the nature of the errors. According to Chenard & Caissie (2008), the average of all individual errors, named bias, can be calculated indicating whether

Table 3 | Statistical characterization of the variables used in the study

	Variable	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation (%)
Database A (52 records)	pH (Sørensen scale)	6.3	8.4	7.5	0.3	4.0
	Conductivity (μS/cm)	152.0	900.0	563.2	180.3	32.0
	Nitrate (mg/dm^3)	1.7	51.0	19.9	14.3	71.8
	Sodium (mg/dm^3)	9.0	80.0	35.0	19.1	54.6
	Potassium (mg/dm^3)	0.41	4.8	2.6	1.3	50.0
Database B (60 records)	pH (Sørensen scale)	6.7	8.7	7.7	0.4	5.2
	Conductivity (μS/cm)	255.0	1,220.0	478.5	186.5	39.0
	Nitrate (mg/dm^3)	1.5	49.0	16.0	13.6	85.0
	Manganese (μg/dm^3)	2.0	30.0	12.6	7.5	51.5

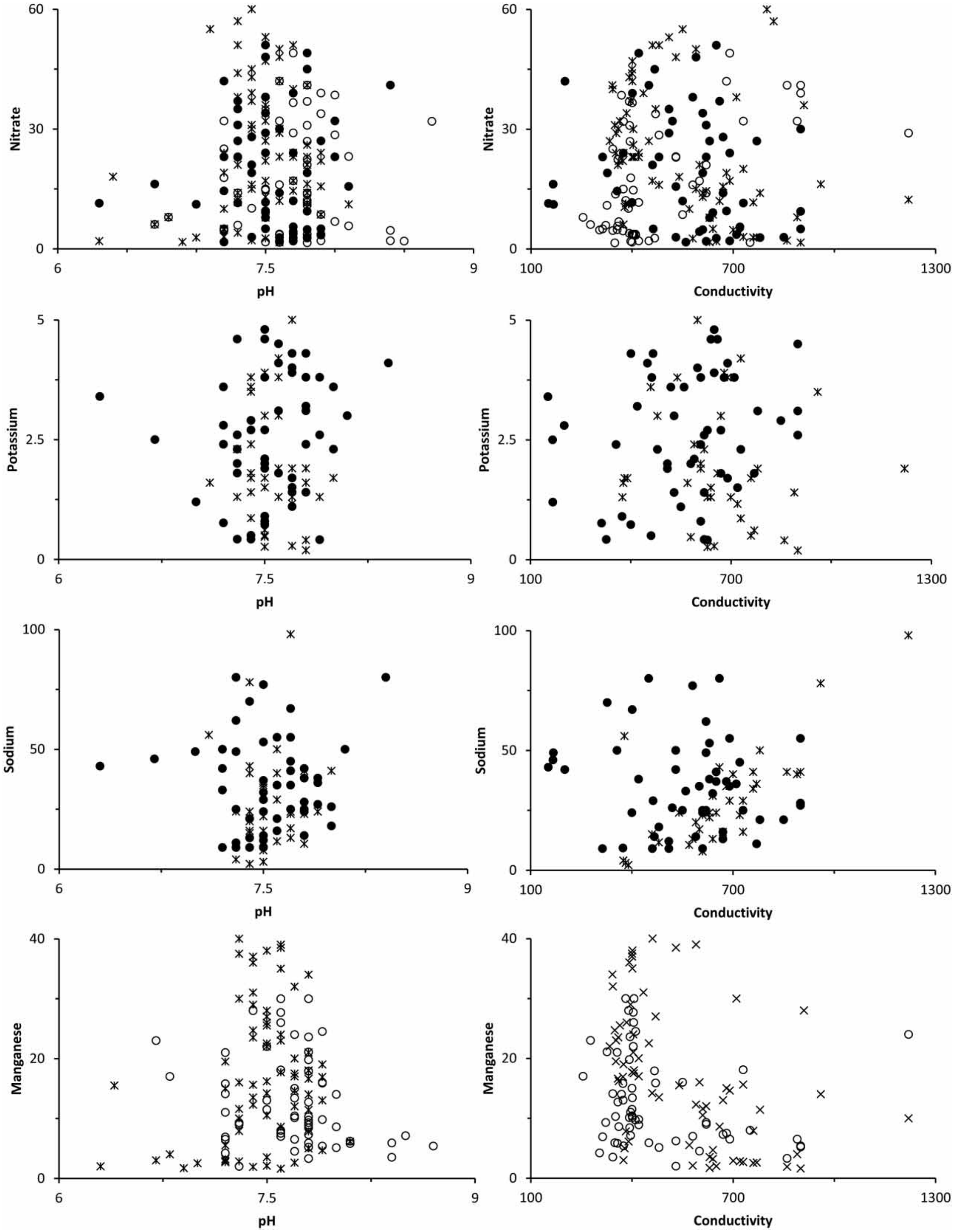

Figure 2 | 2D scatter plots of the dependent variables (nitrate, potassium, sodium and manganese) versus independent variables (pH and conductivity) for database A (•), database B (○) and validation data (*).

the model overestimates or underestimates the output variables. Table 4 presents the values of MAD, MSE and bias for some of the topologies considered.

Concerning the prediction of nitrate, potassium and sodium levels, Table 4 shows that an ANN with a topology of 2-18-14-3 minimizes MAD and MSE, and exhibits a bias value closer to zero for both the training and test sets. Regarding the prediction of nitrate and manganese levels, Table 4 shows that an ANN with a topology of 2-19-10-2 minimizes MAD and MSE, and exhibits a bias value closer to zero for

Table 4 | Mean absolute deviation (MAD), mean squared error (MSE) and bias for some ANN topologies tested

		MAD[a]		**MSE**[a]		**Bias**[a]	
	ANN Topology	**Training set**	**Test set**	**Training set**	**Test set**	**Training set**	**Test set**
Prediction of nitrate, potassium and sodium levels	2-27-3	2.765	3.456	25.823	32.651	−0.654	0.945
	2-17-13-3	1.759	1.853	9.086	9.340	−0.327	−0.755
	2-17-14-3	1.987	1.999	12.281	13.359	0.347	0.929
	2-18-14-3	0.492	0.531	0.573	0.707	−0.0150	0.0561
	2-18-15-3	1.332	1.059	2.023	2.624	−0.292	0.162
	2-19-17-3	0.856	0.983	1.113	0.995	0.107	0.0893
	2-22-18-3	1.849	2.224	12.316	15.416	−0.384	−0.855
	2-22-22-3	2.122	2.363	17.482	16.268	0.356	0.749
Prediction of nitrate and manganese levels	2-21-2	3.152	3.440	19.286	20.315	0.457	0.583
	2-9-5-2	1.563	1.580	3.653	3.760	−0.291	0.495
	2-12-5-2	1.514	1.522	3.422	3.686	−0.342	0.556
	2-14-10-2	0.900	0.921	1.948	2.147	−0.119	−0.335
	2-17-10-2	0.822	0.990	0.951	1.104	0.0468	−0.122
	2-19-10-2	0.372	0.463	0.199	0.274	−0.0113	0.0755
	2-19-17-2	0.524	0.874	0.664	0.767	0.0536	−0.118
	2-24-11-2	2.005	2.719	3.699	3.294	0.184	−0.437

[a] $\mathrm{MAD} = \frac{\sum_{i=1}^{N} |Y_i' - Y_i|}{N}$; $\mathrm{MSE} = \frac{\sum_{i=1}^{N} (Y_i' - Y_i)^2}{N}$; $\mathrm{bias} = \frac{\sum_{i=1}^{N} (Y_i' - Y_i)}{N}$; where Y, Y' and N denote, respectively, an experimental value, a predicted value, and the number of observations.

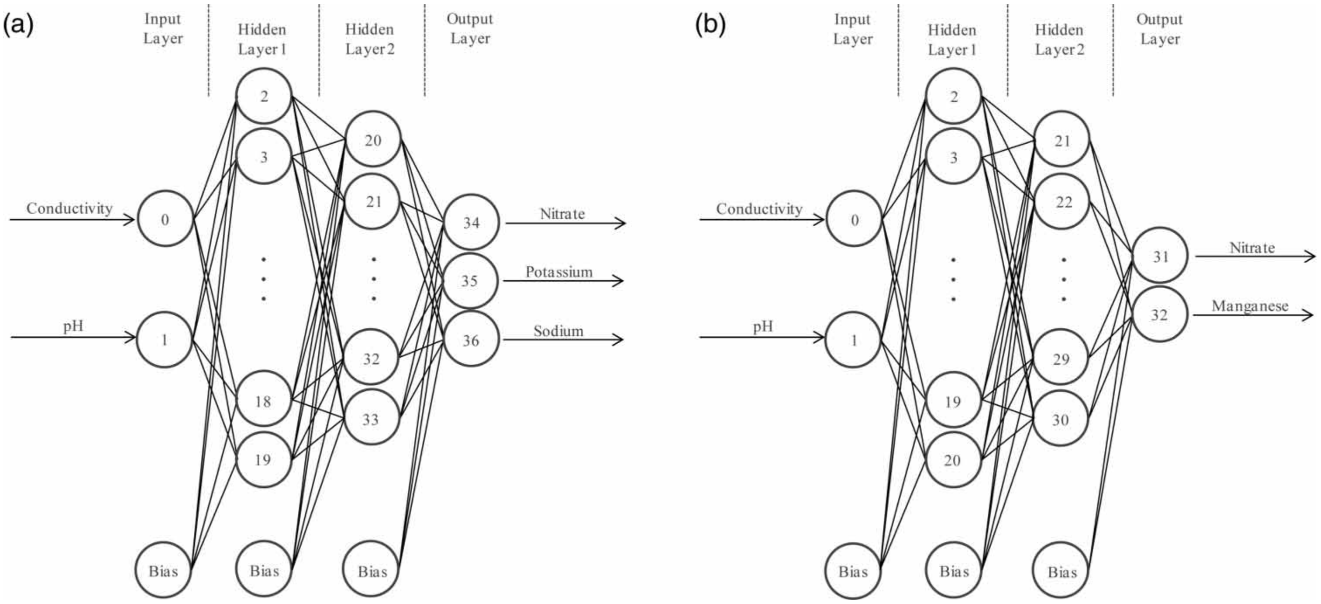

Figure 3 | The ANN structure for modelling nitrate, potassium and sodium (a) and nitrate and manganese levels (b), from conductivity and pH.

Table 5 | Comparison between measured and evaluated responses by the selected ANN models

		Training set			Test set		
		MAD	**MSE**	**Bias**	**MAD**	**MSE**	**Bias**
Model A	Nitrate	0.425	0.308	−0.068	0.298	0.173	−0.005
	Potassium	0.064	5.9×10^{-3}	0.011	0.049	3.4×10^{-3}	0.002
	Sodium	0.986	1.406	0.012	1.246	1.946	0.172
Model B	Nitrate	0.388	0.208	−0.007	0.413	0.249	0.082
	Manganese	0.357	0.191	−0.016	0.513	0.300	0.069

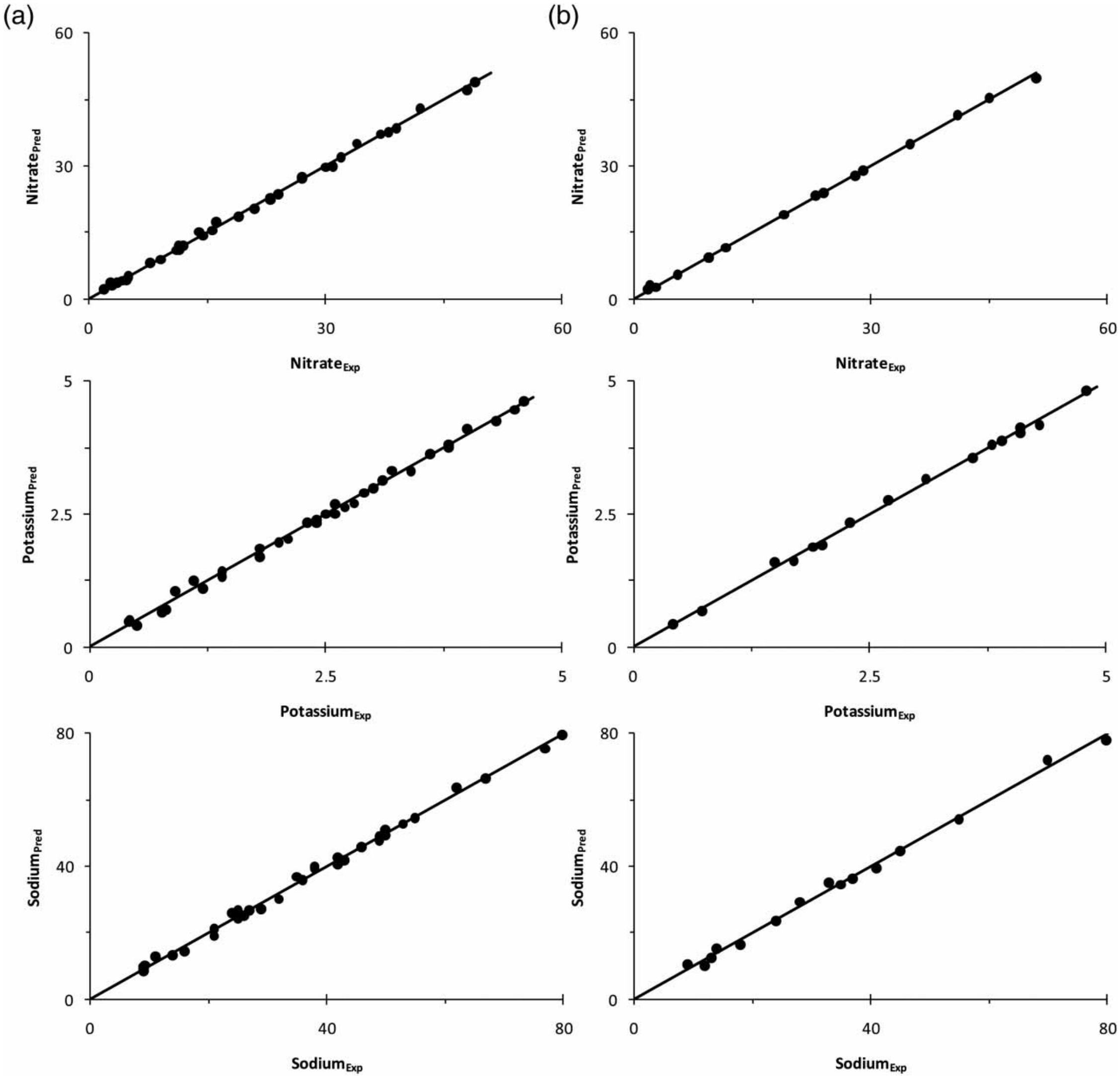

Figure 4 | Plot of the predicted response by the ANN model for nitrate, potassium and sodium versus experimental values for training (a) and test (b) sets.

both data sets. The architecture of the best ANN for modelling the nitrate, potassium and sodium levels in public water supply of the district of Évora (model A) is shown in Figure 3(a), while Figure 3(b) shows the best network architecture for modelling nitrate and manganese levels (model B). Model A consists of an input layer with two nodes, two

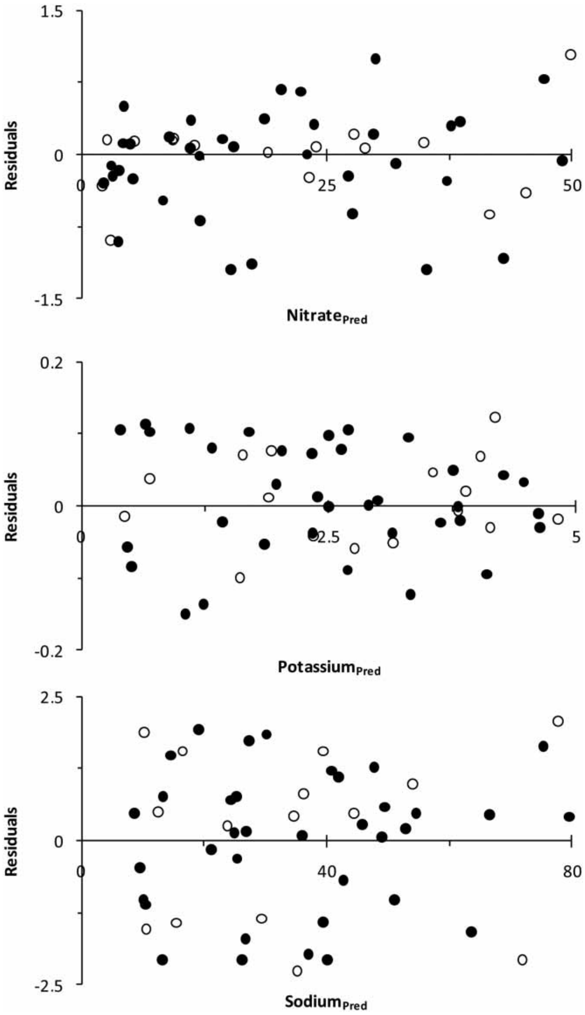

Figure 5 | Plot of the residuals versus the predicted response by ANN model for nitrate, potassium and sodium, for training (•) and test (○) sets.

hidden layers with 18 and 14 nodes, and a three nodes output layer. Model B, in turn, consists of an input layer with two nodes, two hidden layers with 19 and 10 nodes, and a two nodes output layer.

The values of MAD, MSE and bias were computed for the training and test sets, for the two models, models A and B, and are presented in Table 5. Figure 4(a) and (b) shows the mark points on a graph for experimental and predicted values of nitrate, potassium and sodium concentrations for training and test sets. The coefficient of determination (R^2) values ($p < 0.001$) for the training and test sets was 0.998/0.997, 0.996/0.998, and 0.996/0.99 for nitrate, potassium and sodium. The agreement between the experimental and predicted values for the three parameters of water quality (Figure 4(a),(b)), R^2, MAD, MSE and bias suggest a good fit of model A to the data set.

In addition to the above, Figure 5 shows the plots for residuals and predicted values of nitrate, potassium and sodium concentrations for training and test sets. The observed relationship between residuals and predicted values for the three water quality parameters for training and test sets shows complete independence and random distribution. Indeed, the values of the coefficient of determination ($p < 0.001$) are quite small (i.e. R^2 lies between 0.007 and 0.013 for the training set, and between 0.004 and 0.053 for the test set). Figure 5 shows that the marks on the chart are well distributed on both sides of the horizontal line of zero ordinate, corresponding to the correct prediction. Plots of the residuals versus predicted values can be more informative regarding model fitting to a data set. If the residuals appear to behave randomly, it suggests that the model fits the data quite well. On the other hand, if a non-random distribution is obvious in the residuals, the model does not fit the data adequately (McBride 2005).

In terms of the model B for modelling the nitrate and manganese levels in public water supply of the district of Évora from pH and conductivity, the coefficient of determination values ($p < 0.001$) for the training and test sets were 0.9989/0.9961 and 0.9987/0.9958, respectively for nitrate and manganese. Figure 6(a) and (b) shows the plots between experimental and predicted values of nitrate and manganese concentrations for training and test sets. The agreement between the experimental and predicted values for the parameters of water quality, R^2, MAD, MSE and bias (Table 5), suggests a good fit of model B to the data set.

In addition to the above, Figure 7 shows the mark points on a graph for residuals and predicted values of nitrate and manganese concentrations, for training and test sets. The observed relationship between residuals and predicted values for these water quality parameters for training and test sets shows complete independence and random distribution. Indeed, the values of the coefficient of determination ($p < 0.001$) are negligible (i.e. the respective values of R^2 for nitrate and manganese are $0.006/1 \times 10^{-7}$ for the training set and 0.027/0.004 for the test set). Figure 7 shows that the mark points on the graph are well distributed

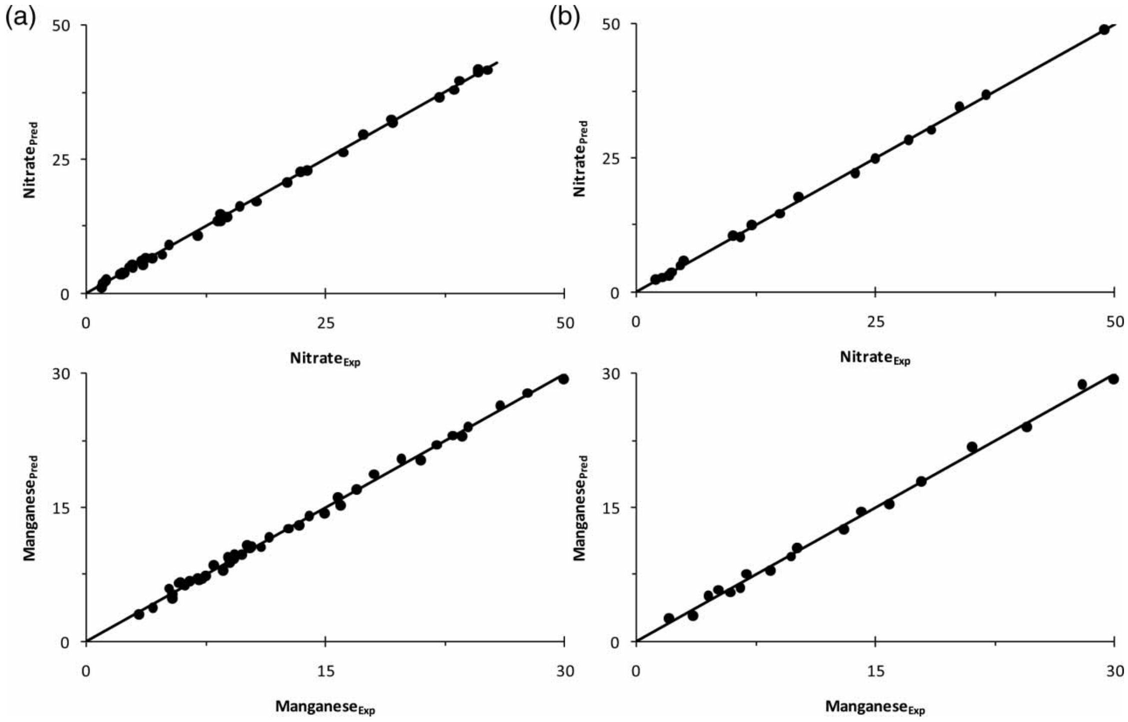

Figure 6 | Plot of predicted response by ANN model for nitrate and manganese versus experimental values for training (a) and test (b) sets.

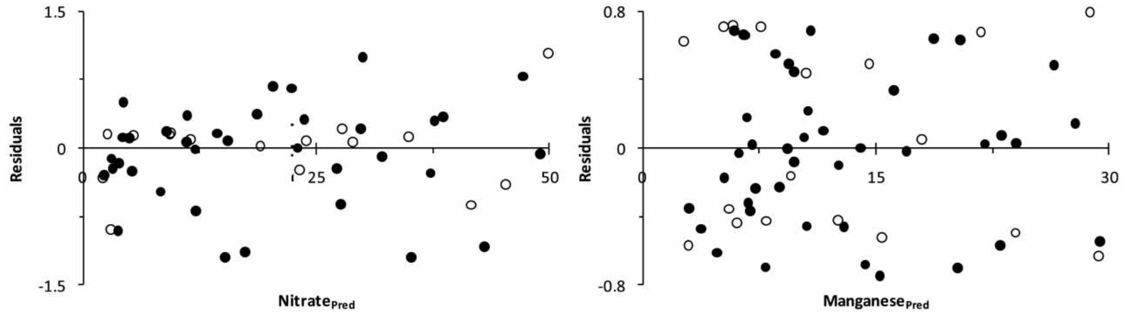

Figure 7 | Plot of the residuals versus the predicted response by the ANN model for nitrate and manganese for training (•) and test (○) sets.

on both sides of the horizontal line of zero ordinate, corresponding to the correct prediction.

Validation of the ANNs models

In order to validate the models, a set of independent data for the year 2010 was used and the values of the output parameters were computed. Figure 8 shows the mark points on a graph for residuals and predicted values of the output parameters for the validation data. The relationships observed for both models show complete independence, random distribution and are well distributed on both sides of the line of correct prediction. The R^2, MAD, MSE and bias were evaluated for both models and are presented in Table 6. Such values are similar to those presented earlier, for model B, namely for the training and test sets. Concerning model A, the results of MSE for validation data increased 30- and 50-fold for nitrate and 290- and 500-fold

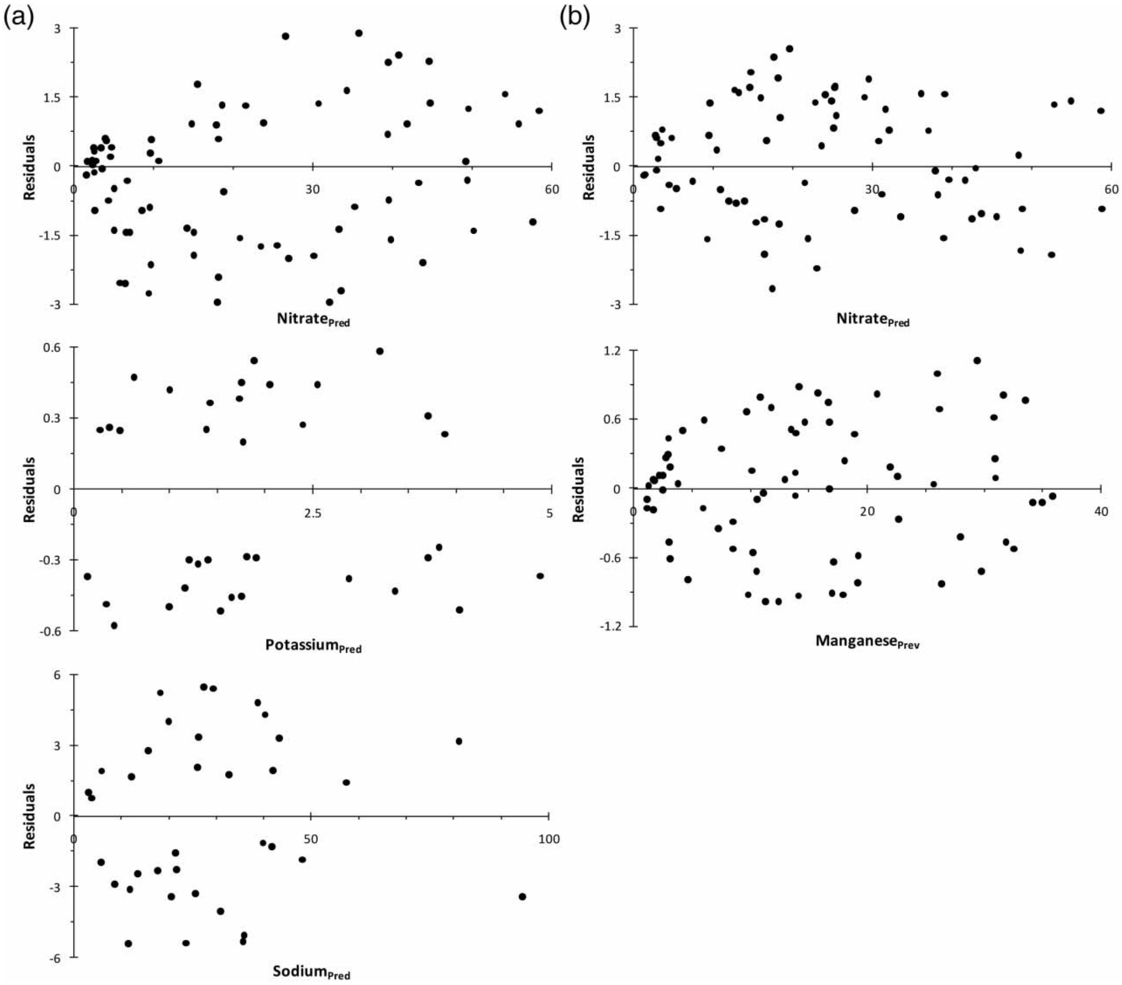

Figure 8 | The mark points on a graph for the residuals versus predicted values of the output parameters for model A (a) and model B (b), in terms of the validation set.

Table 6 | Comparison between measured and evaluated responses by the selected ANNs models for substantive data

		R^2	MAD	MSE	Bias
Model A	Nitrate	0.972	2.407	8.607	−0.096
	Potassium	0.988	1.060	1.712	−0.023
	Sodium	0.970	3.371	11.544	−0.054
Model B	Nitrate	0.989	1.336	2.563	0.084
	Manganese	0.988	1.125	1.890	0.045

for potassium when compared to the training and the test data. Nevertheless, the relationships between the ratio predicted/observed values versus predicted values for the validation data (depicted in Figure 9) seems to converge to the line of correct prediction, meaning that the relative error decreases as the predicted values increase. This trend is quite significant in the process of the quality of public water supplies monitoring, since the predictive capacity of the models should be high when the parameters are close to the maximum allowable, which are set in the national law as 50 mg/dm^3 for nitrate, 200 mg/dm^3 for sodium, and 50 μg/dm^3 for manganese. Potassium is not mentioned in the national law. This parameter does not have an allowable or recommended limit. It is determined for informational purposes since it is analysed simultaneously with the sodium.

The results presented in this study are comparable to those obtained by Yesilnacar *et al.* (2008). These authors employed a 4-25-1 ANN topology for the prediction of nitrate in groundwater, using as input variables pH,

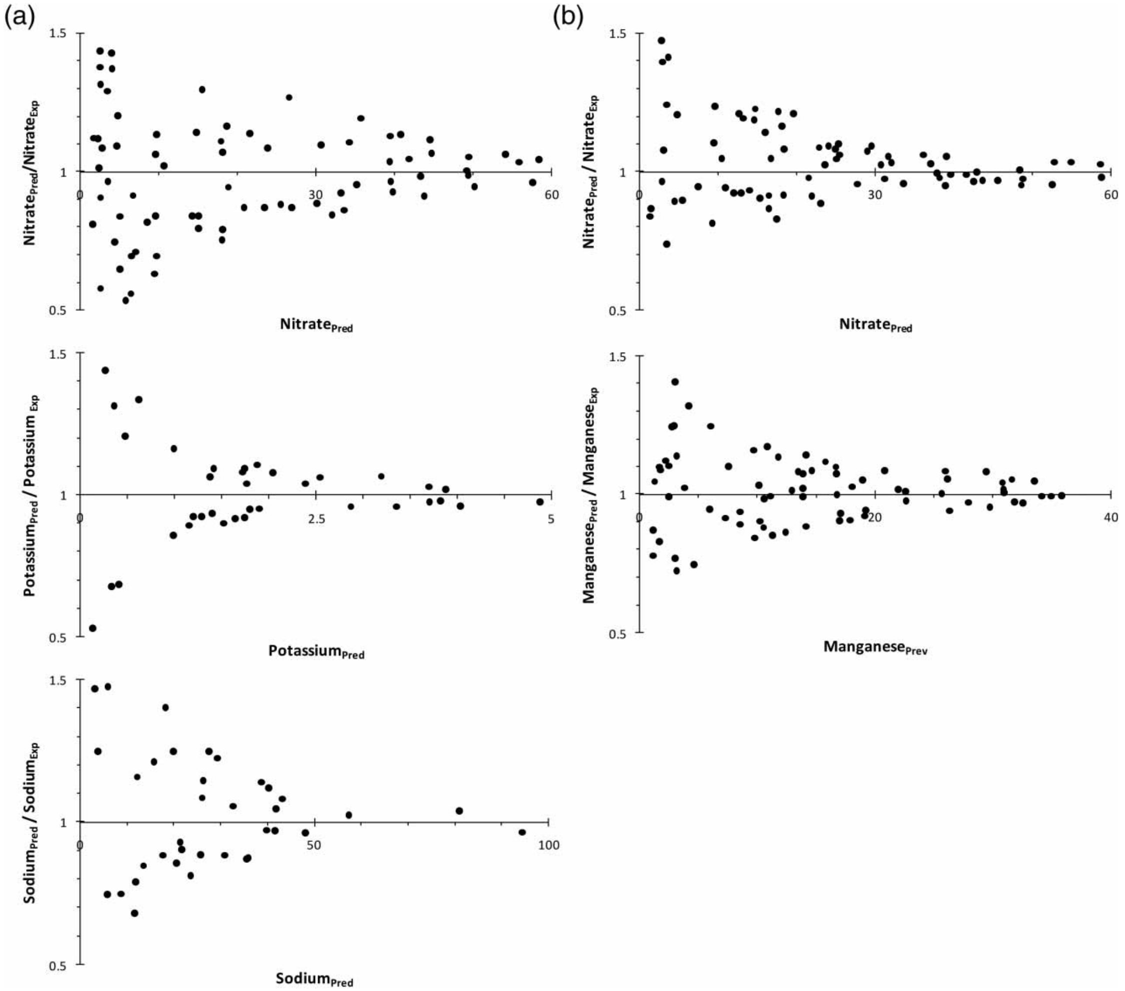

Figure 9 | The mark points on a graph for the ratio between predicted and observed values versus predicted values for the validation set, for model A (a) and model B (b).

temperature, conductivity and groundwater level. The data were taken monthly over 1 year from 24 representative observation wells. The correlation coefficient between the measured and computed nitrate concentrations was 0.93.

Sensitivity analysis of the ANN models

Typically, the efforts in data acquisition will be focused on the more relevant variables for model accuracy, dropping or ignoring those that matter least. Sensitivity analysis is concerned with the model output sensitivity to changes in its input variables. Sensitivity analysis is a simple procedure that is applied after the modelling phase, and analyses the model responses when the inputs are changed. Sensitivity according to variance was used (Kewley *et al.* 2000) to compute the relative importance of the input variables for the selected models. The results are presented in Figure 10, and reveal that the most informative variable is conductivity, for both models, although a relatively higher contribution is observed in model B. These results seem to suggest that the presence of ionic species plays an important role in determining the levels of nitrate, manganese, potassium, and sodium in public water supply, particularly on the contents of nitrate and manganese. These results are in agreement with those obtained by Yesilnacar *et al.* (2008), for the prediction of nitrate in groundwater via a 4-25-1 ANN topology using as input pH, temperature, conductivity and groundwater level. Despite the use of a different approach

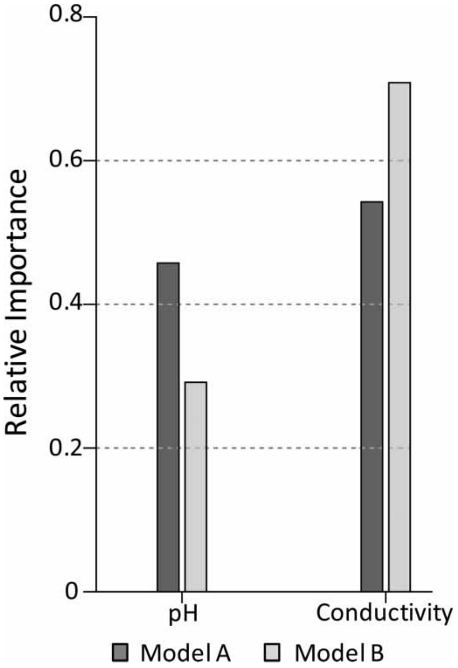

Figure 10 | Relative importance of the input variables for selected ANNs to model nitrate, potassium and sodium (model A), and nitrate and manganese levels (model B), in public water supply of the district of Évora.

to compute the sensitivity, the results show that the most informative variable is the conductivity, followed by the pH.

CONCLUSIONS

In this study a new methodology for problem solving was emphasized. Diverse ANN architectures were developed and evaluated to predict the contents of nitrate, potassium and sodium, and the contents of nitrate and manganese in public water supply of the district of Évora. The data were gathered by the Public Health Laboratory of Évora in the time period from January 2007 to December 2009. The feedforward network with back-propagation learning algorithm was employed. The selected networks performed well in prediction of the output variables based on pH and conductivity for an independent set of data and, therefore, do not show overfitting. The encouraging results obtained in this work show that the ANNs can be very useful as tools to predict water quality parameters and can contribute significantly to the effort that is needed for constant improvement of quality of public water supply.

REFERENCES

Almeida, C., Mendonça, J., Jesus, M. & Gomes, A. 2000 *Aquifer Systems of Portugal*. INAG, Lisbon, Portugal (in Portuguese).

Chenard, J.-F. & Caissie, D. 2008 Stream temperature modelling using artificial neural networks: application on Catamaran Brook. *Hydrol. Proc.* **22**, 3361–3372.

Cheng, B. & Titterington, D. 1994 Neural networks: a review from a statistical perspective. *Stat. Sci.* **9**, 2–30.

Cortez, P., Rocha, M. & Neves, J. 2004 Evolving time series forecasting ARMA models. *J. Heurist.* **10**, 415–429.

Eaton, A., Clesceri, L., Rice, E. & Greenberg, A. (eds) 2005 *Standard Methods for the Examination of Water and Wastewater*. American Public Health Association, USA.

Farmaki, E. G., Thomaidis, N. S. & Efstathiou, C. E. 2010 Artificial neural networks in water analysis: theory and applications. *Int. J. Environ. Analyt. Chem.* **90**, 85–105.

Flood, I. & Kartam, N. 1994 Neural network in civil engineering: I. Principles and understanding. *J. Comput. Civil Eng.* **8**, 131–148.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. 2009 The WEKA Data Mining Software: an update. *SIGKDD Explor.* **11**, 10–18.

Han, J. & Kamber, M. 2006 *Data Mining: Concepts and Techniques*. Morgan Kauffmann Publishers, San Francisco, USA.

Haykin, S. 2008 *Neural Networks and Learning Machines*. Prentice Hall, New Jersey, USA.

Hornik, K., Stinchcombe, M. & White, H. 1989 Multilayer feed-forward networks are universal approximators. *Neural Netw.* **2**, 359–366.

IRAR 2005 *Sampling Procedure for Water to Human Consumption in Public Supply Systems, Recommendation n° 8/2005*. IRAR, Lisbon, Portugal (in Portuguese).

Kewley, R., Embrechts, M. & Breneman, C. 2000 Data strip mining for the virtual design of pharmaceuticals with neural networks. *IEEE Trans. Neural Netw.* **11**, 668–679.

Maier, H., Jain, A., Dandy, G. & Sudheer, K. 2010 Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. *Environ. Model. Softw.* **25**, 891–909.

McBride, G. B. 2005 *Using Statistical Methods for Water Quality Management Issues, Problems and Solutions*. John Wiley & Sons, Hoboken, USA.

Palani, S., Liong, S.-Y. & Tkalich, P. 2008 An ANN application for water quality forecasting. *Mar. Pollut. Bull.* **56**, 1586–1597.

Pinto, A., Fernandes, A. V., Vicente, H. & Neves, J. 2009 Optimizing water treatment systems using artificial intelligence based tools. In: *Water Resources Management V,* *WIT Transactions on Ecology and the Environment, Vol. 125* (C. A. Brebbia & V. Popov, eds). WIT Press, Southampton, UK, pp. 185–194.

Rumelhart, D., Hinton, G. & Williams, R. 1986 Learning internal representation by error propagation. In: *Parallel Distributed Processing, Vol. 1: Foundations* (D. E. Rumelhart & J. L. McCleland, eds). MIT Press, Massachusetts, USA, pp. 318–362.

Santos, M. F., Cortez, P., Quintela, H., Neves, J., Vicente, H. & Arteiro, J. 2005 Ecological mining – a case study on dam water quality. In: *Data Mining VI – Data Mining, Text Mining and their Business Applications, WIT Transactions of Information and Comunication Technologies, Vol. 35* (A. Zanasi, C. A. Brebbia & N. F. F. Ebecken, eds). WIT Press, Southampton, UK, pp. 523–531.

Singh, K., Basant, A., Malik, A. & Jain, G. 2009 Artificial neural network modeling of the river water quality – a case study. *Ecol. Model.* **220**, 888–895.

Souza, J., Matwin, S. & Japkowicz, N. 2002 Evaluating data mining models: a pattern language. In: *Proceedings of 9th Conference on Pattern Language of Programs*. Illinois, USA.

Torgo, L. 1999 Inductive Learning of Tree-Based Regression Models. PhD, University of Oporto, Oporto, Portugal.

West, D. & Dellana, S. 2011 An empirical analysis of neural network memory structures for basin water quality forecasting. *Int. J. Forecast.* **27**, 777–803.

Yesilnacar, M. I., Sahinkaya, E., Naz, M. & Ozkaya, B. 2008 Neural network prediction of nitrate in groundwater of Harran Plain, Turkey. *Environ. Geol.* **56**, 19–25.

First received 17 February 2012; accepted in revised form 7 September 2012