Prediction of the optimal dose of coagulant for various potable water treatment processes through artificial neural network

M. Hassen Baouab and Semia Cherif

ABSTRACT

To overcome classical jar test limits of water treatment plants and offer substantial savings of time and money for operators, artificial neural network technique is applied in this study to large databases of three treatment plants with different processes in order to build models to predict the optimal dose of coagulant. Pre-modeling techniques, like data scaling and training database choice, are used to guarantee models with the lowest errors. Two models are then selected, with turbidity, conductivity, and pH as inputs for both raw and treated water. The first model, L45-MOD, is specific to raw water with less than 45.5 NTU turbidity, or else the second model ATP-MOD would be adopted. Compared to truly injected coagulant doses and to previous models, the selected models have good performances when tested on various databases: a correlation coefficient higher than 0.8, a mean absolute error of 5.47 g/m^3 for the first model and 5.69 g/m^3 for the second model. The strength of this study is the ability of the models to be extrapolated and easily adopted by other treatment plants whatever the process used.

Key words | artificial neural network, coagulant, drinking treatment plants, potable water, prediction

M. Hassen Baouab
Semia Cherif (corresponding author)
UR Chimie des Matériaux et de l'Environnement
UR11ES25, ISSBAT,
Université de Tunis El Manar,
Tunis,
Tunisia
E-mail: *semiacherif@yahoo.fr*

INTRODUCTION

Coagulation-flocculation is a conventional technique and a critical process of the drinking water treatment (Lamrini *et al.* 2005). In water treatment plants (WTP), the process control is generally accomplished by examining the quality of the produced water then adjusting the processes according to the operator's own experience (Wu & Lo 2008). As well, the coagulant dose needed for coagulation is determined empirically through laboratory jar test, where a single test may take at least 1 hour to be performed; during periods of fast variations of water characteristics (e.g., during floods) it is impossible to have a real-time response. Moreover, this conventional technique, due to manual intervention, can lead to excessive or insufficient non-adequate coagulant doses (Lamrini *et al.* 2005). The SCD method (Streaming Current Detector) is another conventional method to determine the adequate coagulant dose. Its output is correlated to the measured zeta potential, responsible for the water particle dispersion and cohesion. As well as for the jar test, the SCD has limits, such as its operation cost and a limited efficiency when the raw water has a pH of 8 (Lamrini *et al.* 2005). Besides these conventional techniques (jar tests and SCD), there is no single measure to assess the overall quality of the raw or treated water and thus the efficiency of the purification cannot be well controlled or optimized. Therefore, there is a crucial need for a fast and real-time response technique that can deliver the optimal coagulant dose and help the optimization of the treatment process.

Optimizing the treatment process represents a current challenge due to the importance of the treated water quality,

doi: 10.2166/hydro.2018.014

the consumer satisfaction, the complexity of the treatment process (Li *et al.* 2017), and the impact of emerging factors such as climate change and socio-political disturbances (Baouab & Cherif 2015, 2017a). Optimization needs a powerful, fast, and reliable method for determining the optimal dose of coagulant to add in the treatment process. In this context, an artificial intelligence approach through artificial neural networks (ANNs) can lead to a robust model that meets the operator's needs: time and money saving with a guarantee of a clear water respecting standards. ANNs are a family of statistical learning algorithms that learn the relationships among the variables through successive training, having the ability to learn patterns that are not linearly separable and concepts dealing with uncertainty, noise, and random events (Dębska & Guzowska-Świder 2011).

Various applications of ANN have been reported in the drinking water treatment industry: prediction of the drinking water consumption (Guhl & Brémond 2000; Koffi *et al.* 2012), leak detection and calibration (Vítkovský *et al.* 2000), predicting membrane fouling (Delgrange-Vincent *et al.* 2000; Shetty & Chellam 2003), adequate chlorine dosage for potable water disinfection (Rodriguez & Serodes 1996), and prediction of treated water quality parameters (Kabsch-Korbutowicz & Kutylowska 2011; Rak 2013). Specifically, many researchers have focused their studies on prediction of coagulant dosing: Bazer-Bachi *et al.* (1990), Baba *et al.* (1990), Gagnon *et al.* (1997), Van Leeuwen *et al.* (1999), Valentin (2000), Maier *et al.* (2004), Lamrini *et al.* (2005), Wu & Lo (2008), Heddam *et al.* (2012), and recently, Kumar *et al.* (2013) who developed a model to control the coagulant dosing pump.

All the cited studies predicting the optimal dose of coagulant are limited to the direct use of ANN with no preliminary data treatment or input choice through factor reduction. Even when factor reduction is used, it is applied to one type of water treatment process and only on raw water parameters, besides the fact that the vast majority developed models adapted for a specific water treatment plant. Van Leeuwen *et al.* (1999) and Maier *et al.* (2004) developed models for different case studies, but the multiplicity concerned only water sources (raw water).

This article stands out due to the development of ANN models to predict the optimal dose of coagulant for three types of treatment processes: static settlement tank, lamella separator, and pulsator settlement tank. Unlike in many other studies, the developed models include as inputs, for both raw and treated water, parameters that can be instantly measured. As well, a large part is allocated to the generalization of the developed models with the aim of being an easy-to-use tool whatever the water treatment plant process.

STUDIED PLANTS

Conventional potable water treatment processes may vary slightly, depending on the technology of the plant, but usually include coagulation, flocculation, sedimentation, and filtration.

During coagulation and flocculation, chemicals (coagulant and flocculant) are added to raw water. This causes the tiny particles of dirt in the water to stick together and coagulate. As these particles become larger, they are called flocs, which is the flocculation. The next step is the sedimentation wherein the heavy flocs are allowed to settle.

Three WTP, situated in Tunis and using the main processes cited above, with different sedimentation processes, are studied. The first water treatment plant (TP1) has a static settlement tank, the second one (TP2) has a pulsator settlement tank, and the third (TP3) has a lamella separator. TP2 and TP3 have the same source of raw water which is different from that of TP1.

DATA AND METHODS

Data

Large databases are used in this study: a 935×7 matrix for TP1, $1{,}445 \times 7$ matrix for TP2, and $1{,}469 \times 7$ matrix for TP3. In each of these matrixes, columns represent the instrumental analysis values: turbidity, conductivity, and pH of raw water; turbidity, conductivity, and pH of treated water, and the coagulant dose. Rows represent the day of sampling, from 2007 to 2012.

The data were collected for a long time (almost daily from 2007 to 2012) and from different WTP to give a real

significance to the mathematical computation, so that the obtained results could eventually be extrapolated for applications to industrial treatment plant management.

To guarantee that the ANN model generates the optimal dose of coagulant, the studied databases include only samples with treated water parameters meeting standards. The Tunisian standard NT09.14 is adopted as a reference. Its values meet those recommended by the World Health Organization and many other standards like those of the European Union, Germany, United Kingdom, and France.

In order to ensure that all the variables receive equal attention during the training process, the database should be scaled. However, when the functions in the output layer are linear, scaling is not strictly required but recommended (Maier & Dandy 2000). In our study, the output layer functions are linear, and consequently, the data are scaled before the analysis to have results easily comparable from one model to another. This pre-processing technique scales all numeric variables in the range [0, 1] by performing Equation (1):

$$Xij\text{scaled} = \frac{(Xij - min(Xj))}{(max(Xj) - min(Xj))} \quad (1)$$

where Xij is the data on the ith row and jth column of the database matrix and the $Xijscaled$ is the corresponding scaled value, $min(Xj)$ and $(maxXj)$ are, respectively, the lowest and highest X value in the jth column.

Method

Data reduction approaches by principal component analysis (PCA) and hierarchical clustering (HCL) were applied to the same databases as the ones used to develop the ANN model in the present paper, for all the treatment plants (Baouab & Cherif 2017b). It determined which of the multiple parameters (turbidity, salinity, conductivity, pH, M alkalinity, water hardness, calcium, magnesium, chlorine, and organic matter) could represent significantly the variation of water quality during the treatment process. The significant parameters are identified as turbidity, conductivity, and pH, all for raw and treated water. These parameters are used as inputs for modeling by ANN and by multivariable regression (MRG) to obtain the optimal dose of coagulant as output.

To examine inputs' distribution, the Shapiro–Wilk test is adopted. This test, commonly used, is based on the W coefficient calculated in this study by the software R. The higher is W (closer to 1), the higher is the compatibility with the normal distribution (Rakotomalala 2008).

For modeling, these main steps are followed: the division and pre-processing of the available data, the choice of the training data to ensure model generalization, inputs, network architecture, training, testing, and model validation.

The modeling starts by the division of the complete data set into a training and an evaluation (test) set: samples of databases' rows are assigned randomly to a training set, consisting of 70% of all the samples. The remaining 30% composed the test set. Thus, after testing different training databases, the one that can provide a model with low errors for all the WTP test databases is chosen in order to have a general model useful for any type of water treatment process. Sigmoid function is used, during training, as the activation function. This function restricts the output interval to [0–1] (Dreyfus 2005).

There are a wide variety of ANN in the literature that are known to be able to successfully represent complex functions in various fields. The multilayer perceptron is chosen because it is one of the most commonly used for modeling (Lamrini *et al.* 2005) due to its success in the prediction of water resources variables (Maier & Dandy 2000). It consists of simple processing elements arranged in layers (input layer, hidden layers, and output layer). Each element or node takes its input from the weighted sum of the outputs of the element of the previous layer. This input is then used in a nonlinear function, often called the activation function, to form the element's output (Lamrini *et al.* 2005).

In order to avoid over-fitting and to choose the network architecture represented by the number of hidden layers and the number of nodes in each, cross-validation (CV) technique is used with different combinations of hidden layers and node numbers. CV starts by splitting the training data created above, once or several times (Arlot & Celisse 2010). Each time, the ANN model is built for a specific number of nodes. Errors for each model architecture are calculated and the one with fewer errors is selected. These

steps (splitting, model building, error calculation, and model selection) are repeated every time the number of nodes is changed.

The CV results represented by the models with various numbers of nodes, one model for each number of node and associated errors, are summarized and compared with the aim to choose the optimal model that has the best accuracy in the predictive output. The choice is based on the assessment of graphical error summarizing diagram (Taylor diagram) and on the assessment of calculated error criteria. The most commonly used criteria to assess the performance of the models are the mean absolute error (MAE) function (Equation (2)), the root mean squared error (RMSE), and the MSE. However, outliers in training data can affect the errors and their distribution. In order to overcome this problem, the mean log squared error (MLSE) proposed by Liano (1996) as a robust error criterion against outliers is used and calculated by Equation (3) (Zhao & Xu 2005), instead of MSE and RMSE. MAE and MLSE are the adopted error criteria to be calculated for evaluating the best model:

$$MAE = \frac{\sum_{i=1}^{n} ||(Yi - \hat{Y}i)||}{n-2} \quad (2)$$

$$MLSE = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \frac{1}{2}||(Yi - \hat{Y}i)||^2\right) \quad (3)$$

where Yi is the observed value, $\hat{Y}i$ is the predicted value corresponding to Yi, and n is the number of observations or samples.

In combination with calculated error criteria, bivariate correlation analysis is used to examine and explain the relationships between the pairs of variables. The numerical level of the relationships is represented by the correlation coefficients that vary between -1 and $+1$ according to the statistical significance of the estimated correlation. In general, a correlation coefficient greater than 0.50 is statistically significant at a 95% confidence level (Astel *et al.* 2006).

The data analyses are performed using the software R, 'Rcmdr' for the calculation of the Shapiro–Wilk coefficient W, 'lm' function for the MRG, and 'nnet' package for the ANN analysis.

RESULTS AND DISCUSSION

Inputs and data pre-processing

The selection of appropriate inputs is extremely important when building a model (Faraway & Chatfield 1998). The application of PCA and HCL on the present databases showed, in a previous study, that wherein processes are different (static settlement tank, lamella separator, and pulsator settlement tank) at different water treatment stages (raw water and treated water), the same result is always obtained: the parameters are divided into three clear groups. It appears that salinity, conductivity, water hardness, magnesium, calcium, and chlorides are highly explained by the first principal component (PC1) for raw and treated water (with all contributions higher than 0.899), turbidity and organic matter are well explained by PC2 for raw water and by PC3 for treated water (with contributions higher than 0.700). pH and M alkalinity are well correlated to PC3 for raw water and to PC2 for treated water (with contributions higher than 0.730). The results obtained by PCA are consistent with those of HCL whatever the treatment process and its stage. Based on these results, the assessment of the water quality through water treatment process can be done by the evaluation of only three significant parameters, i.e., conductivity, turbidity, and pH (Baouab & Cherif 2017b). They can be easily and quickly measured *in situ* and in the laboratory. These parameters are representative of potable water quality during the treatment process, for raw and treated water, whatever the conducted process: static settlement tank, lamella separator, or even pulsator settlement tank.

Turbidity, pH, and conductivity, the extracted parameters by PCA and HCL, will be considered as the inputs. Thus, the model to be built has six inputs: turbidity, pH, and conductivity, for raw and for treated water.

The six selected inputs that are used for training cover a wide range of values (Table 1) meeting the local and international standards. In the local standard NT09-14, for treated water, the maximum acceptable (or allowable) value for turbidity is 2 NTU, the conductivity value must be between 300 and 2,500 μS, and the pH between 6.5 and 8.5. The data are issued from three potable treatment plants (TP): TP1 with a settlement tank, TP2 with a pulsator settlement tank, and TP3 with a lamella separator. For raw

Table 1 | Raw and treated water parameter extrema

	Raw water			Treated water		
	Turbidity (NTU)	Conductivity (μS)	pH	Turbidity (NTU)	Conductivity (μS)	pH
TP1						
Minimum	2.0	659.0	7.2	0.2	604.0	6.8
Maximum	39.4	3,100.0	8.6	2.0	2,497.2	8.2
TP2						
Minimum	2.2	363.0	7.3	0.1	381.0	6.7
Maximum	174.0	2,980.0	8.5	2.0	2,500.0	8.3
TP3						
Minimum	2.2	178.0	7.3	0.2	385.0	6.6
Maximum	318.0	3,070.0	8.5	2.0	2,500.0	8.2
Merged						
Minimum	2.0	178.0	7.2	0.1	381.0	6.6
Maximum	318.0	3,100.0	8.6	2.0	2,500.0	8.3

water turbidity (Table 1): (i) TP1 database values vary within a narrow range from 2.0 to 39.4 NTU; (ii) TP2 database values vary between 2.2 and 174 NTU; (iii) TP3 database values vary from 2.2 to 318.0 NTU; and (iv) all three databases merged together cover a larger database for all the held parameters. Concerning treated water, in most of the samples, the turbidity is below 2.0 NTU, the conductivity maximum value equals 2,500 μS, the pH maximum value equals 8.3 and its minimum equals 6.6. Only samples with inputs in accordance with the local and international standards are included in these databases.

The relation between input indices and forecasts was analyzed. Hair *et al.* (1995), Tabachnick & Fidell (2007), and Williams *et al.* (2012) recommend a coefficient correlation higher than 0.3 to confirm correlation between variables. The Pearson critical value for a significance level of $\alpha = 0.01$ and samples higher than 500 (Sockloff & Edney 1972) is equal to 0.115. Concerning our databases that have more than 900 samples, calculated Pearson coefficients showed that a correlation exists between input indices and the dose of coagulant (all the calculated coefficients are superior to |0.115|). Moreover, Bartlett's test applied on the studied databases was less than 0.05 indicating that there are significant relationships among variables.

In fact, the Pearson coefficient correlation is not so high, for this reason, neural network was adopted to develop models based on nonlinear correlated variables (Dębska & Guzowska-Świder 2011). Given the complexity of the phenomena (chemical, physical, and biological) involved in drinking water treatment processes and, more specifically, the flocculation coagulation process, it is generally very difficult to quantify the interactions between the parameters which are strongly nonlinear (Valentin 2000).

In the modeling process, the normality of the input distribution is a dilemma. Indeed, in most traditional statistical models, the data have to be normally distributed before the model coefficients can be estimated efficiently (Maier & Dandy 2000; Takahama *et al.* 2004). Some other researchers in the literature explain that ANNs overcome this problem, as the probability distribution of the input data does not have to be known (Maier & Dandy 2000). Nevertheless, in this study, all the inputs were normally distributed: the W coefficient for the majority of the inputs was higher than 0.886. This distribution can be explained by the concentration of the inputs' concentrations around an average value with little variation: (i) for raw water inputs, a seasonal variation exists with punctual water quality change and (ii) for treated water inputs, they vary in their potable water standard range.

The databases used for training are normalized by applying Equation (2). In this way, the built models can be compared according to their errors independently from the

range of the used inputs as long as these inputs are between 0 and 1.

ANN models architecture and database choice, training, and validation

Modeling by ANN depends on the choice of the optimal number of hidden layers and nodes on each of the layers. The use of one hidden layer is common (Hornik *et al.* 1989; Li 1996; Lamrini *et al.* 2005) and it has been proven that a network with one hidden layer can approximate any continuous function (Hornik *et al.* 1989). Therefore, the use of one hidden layer would be sufficient to build the ANN models. We can also consider that the maximum number of nodes equals $2n+1$ in the hidden layer with n the number of model inputs (Hecht-Nielsen 1989; Maier *et al.* 2004). Here, since the number of inputs is equal to 6, the maximum number of nodes is 13. CV is used to build ANN models with different architectures: one hidden layer with a different number of nodes each time.

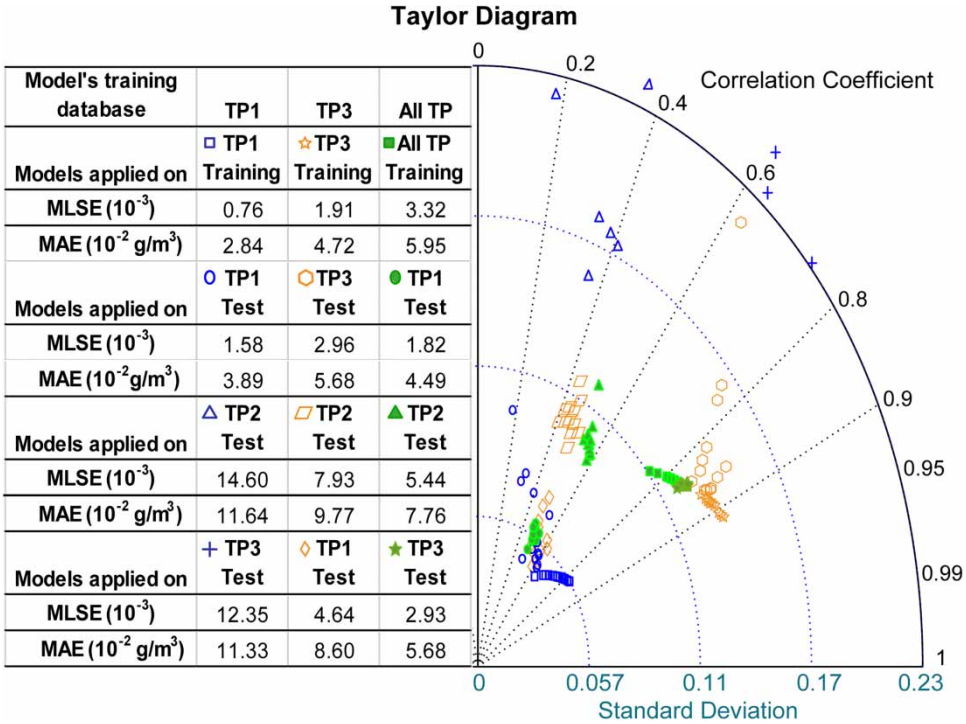
In order to improve the generalization of the built model, three training databases with different ranges are used to select the one that leads to a low error for all the available data of the different plants. (i) TP1 is the first database used: 70% of it (655 samples) is used for training and the remaining 280 samples (TP1 test) are used for the test. The TP2 and TP3 databases, which are not included in training, are used for testing and validating the obtained models. (ii) The second training database concerns 70% of the TP3 database corresponding to 1,028 samples chosen randomly. The remaining 441 observations represent the test database (TP3 test). As before, the remaining databases, this time TP1 and TP2, are used to test the performance of the developed models. (iii) The third training database is the one that mixes the databases of the three WTP. This training database includes 2,710 samples (All TP training) and the tests databases will be from the parts of TP1 (test TP1), TP2 (test TP2), and TP3 (test TP2) databases that are not used for the training. TP2 database was not used in training due to the similarities with TP3: raw water inputs of TP2 and TP3 are similar, and treated water inputs vary in the same range (complying with the standards). However, TP2 database was used at test stages because it is useful to use data of a different water treatment process and useful to use the maximum of samples in tests in order to check models' validation and generalization.

The division of the database ensures that the divided data sets contain all of the patterns present in the data and are thus representative of the same statistical population. Furthermore, division of the database into training and testing subsets ensures that over-fitting does not occur and that the validation data are not used as part of the model development process (Maier & Dandy 2000). After that, and for each built model (13 models by training database, each with a unique architecture, each with a different number of nodes in the hidden layer, from 1 to 13 nodes), the performances are tested by the application of the model on the remaining databases issued from various WTP to check for its capacity of extrapolation.

To assess performances of the models, their error criteria, represented by MLSE, MAE, standard deviation (SD) and correlation coefficient, are calculated and summarized in Figure 1. The figure, besides error criteria, includes the Taylor diagram, which is a diagram that can provide a concise statistical summary of how well patterns match each other in terms of correlations and the ratio of their variances or standard deviations (Taylor 2001).

Training TP1 and test TP1 points (corresponding to models with different architectures based on TP1 training and test databases) are close to the center of the Taylor diagram (Figure 1), an indication that TP1 models have a low SD when applied to their own training or test databases. However, points of TP1 models tested on other TP data (TP2 test and TP3 test) are far from the center, due to their high SD and their low correlation coefficient compared to the other models (Figure 1). All TP models have the highest MLSE and MAE with, respectively, 3.32×10^{-3} and 5.95×10^{-2} g/m^3 when applied to their training databases. However, when the models are tested on the remaining databases, the order of performance changes: the MAE for 'All TP models' reaches 7.76×10^{-2} g/m^3 which is lower than the MAE reached by TP1 models (11.64×10^{-2} g/m^3) and TP3 models (9.77×10^{-2} g/m^3). As well, the MLSE reached by 'All TP models' (5.44×10^{-3}) is still lower than the ones reached by the other models (14.60×10^{-3} for TP1 models and 7.93×10^{-3} for TP3 models).

If the same approach is adopted with the original non-normalized databases, the MAE errors for TP1 training are



Model's training database	TP1	TP3	All TP
Models applied on	TP1 Training	TP3 Training	All TP Training
MLSE (10^{-3})	0.76	1.91	3.32
MAE (10^{-2} g/m^3)	2.84	4.72	5.95
Models applied on	TP1 Test	TP3 Test	TP1 Test
MLSE (10^{-3})	1.58	2.96	1.82
MAE (10^{-2}g/m^3)	3.89	5.68	4.49
Models applied on	TP2 Test	TP2 Test	TP2 Test
MLSE (10^{-3})	14.60	7.93	5.44
MAE (10^{-2} g/m^3)	11.64	9.77	7.76
Models applied on	TP3 Test	TP1 Test	TP3 Test
MLSE (10^{-3})	12.35	4.64	2.93
MAE (10^{-2} g/m^3)	11.33	8.60	5.68

Figure 1 | Taylor diagram summarizing errors (the lowest MAE and MLSE, the SD and the correlation coefficient) generated by different models with different architecture.

the lowest with 4.20 g/m^3. However, when the models are applied to tests, the All TP models give the lowest MAE (7.11 g/m^3 when applied to TP2 test and 5.16 g/m^3 when applied to TP3 test). These results are due to the small range of the TP1 database and the large range of the All TP database and due to the dependency of the MAE on the range limits of the databases when they are not normalized.

Due to the results above, the rest of the development of ANN models is based on the merged TP databases (All TP). In this way, the first step is achieved by the choice of the training database that gives the lowest errors.

The obtained results demonstrate that a database with a narrow parameters' range gives better results in some cases, particularly when it is applied to its own training database and so to a database with the same range. Thus, a narrower range database is extracted from the All TP database and the developed models with All TP as a training database are compared to a model built from the extracted database. The selection of this database is based on a clustering classification (HCL) of raw water turbidity due to the importance of this parameter in the water treatment process. The HCL is applied according to the Ward method, which indicates that the distance between two clusters is related to the increase in the sum of squares when merged. Calculation start with a single variable in its cluster. Clusters then are merged based on the calculation of the distance between them by the Euclidian function. Ward's method keeps this merging growth as small as possible (Baouab & Cherif 2017b). Therefore, HCL applied to raw water turbidity of the All TP database classifies the samples into two groups: the first includes data with a raw water turbidity lower than 45.5 NTU (L45 database) and the second group includes data with a raw water turbidity higher than 45.5 NTU (H45 database).

Three models are built and compared. They are built according to the training database (All TP database, L45 database, and H45 database). Non-normalized databases are used to choose the model that has the best accuracy for the coagulant dose for all the tested databases and that has the ability to be extrapolated. The optimum ANN architecture for each model is selected based on the error of the test set (MAE and SD shown in Figure 2).

Both All TP models and L45 models generate low errors (respectively an average MAE for all tested architectures of

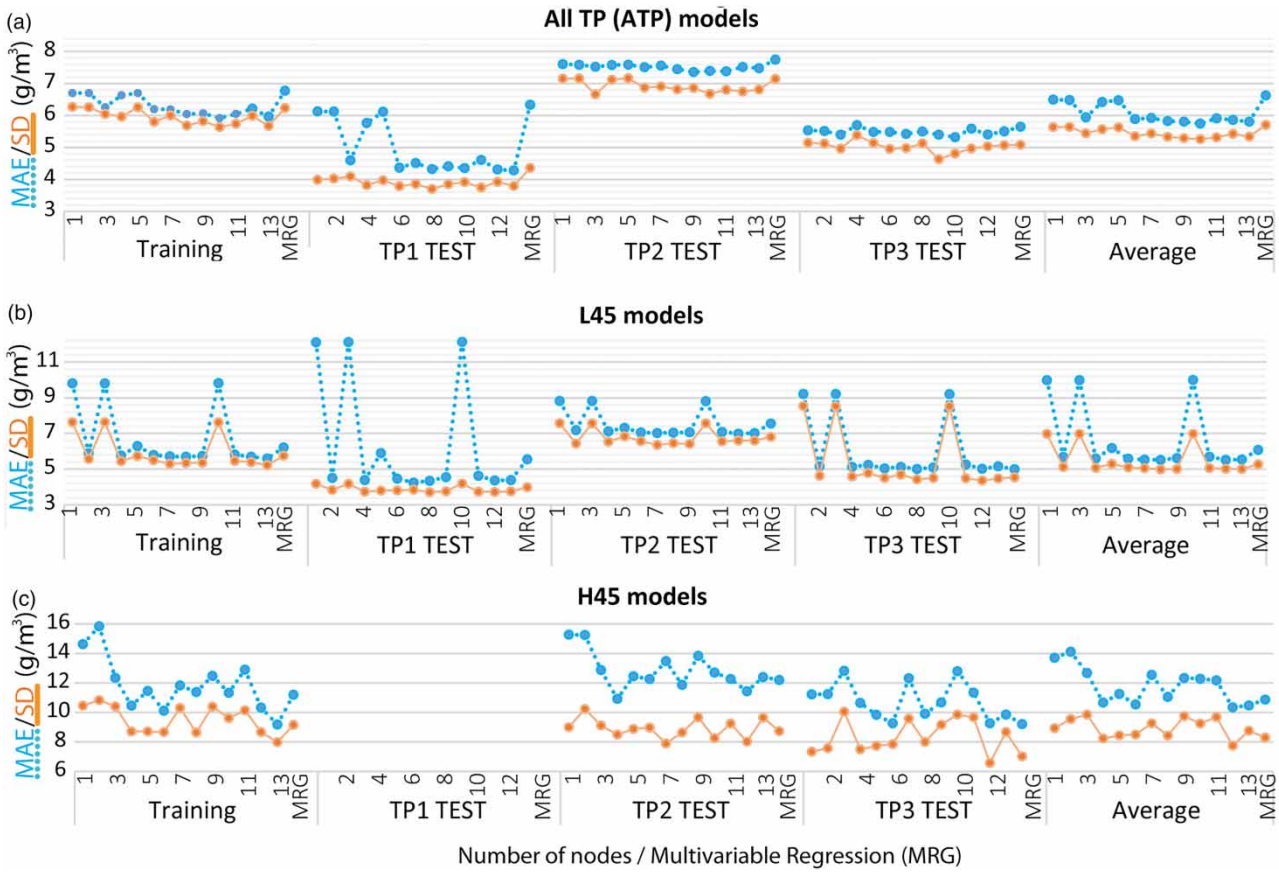


Figure 2 | MAE and SD (g/m^3) generated by ANN models and MRG models: (a) All TP as training database; (b) L45 as training database (raw water turbidity ≤45.5 NTU); (c) H45 as training database (raw water turbidity >45.5 NTU).

6.01 g/m^3 and 6.58 g/m^3) compared to H45 models that reach an MAE of 7.70 g/m^3 for all tested architectures. For All TP models, the one with ten nodes in the hidden layer shows the lowest MAE (4.35 g/m^3) and SD (3.92 g/m^3). Concerning L45 models, the one with eight nodes in the hidden layer shows the lowest MAE (4.35 g/m^3) and SD (3.71 g/m^3). Furthermore, these two architectures (ten nodes for All ATP and eight nodes for L-45) give better results than the models built by MRG: as shown in Figure 2, the average MAE for tested databases is 5.69 g/m^3 for All TP model (with ten nodes) and 5.47 g/m^3 for L45 model (with eight nodes), while the MRG model generates an average MAE for the tested databases of about 6.58 g/m^3 (All TP database as a training database) and 6.03 g/m^3 (L-45 database as a training database).

Thus, one hidden layer with ten nodes is the best architecture for All TP as a training database and the one with eight nodes in the hidden layer is the best architecture for a database with raw water turbidity lower than 45.5 NTU. Models built with a database including raw water turbidity higher than 45.5 NTU are eliminated due to the high MAE generated. The model based on the All TP database will be called ATP-MOD and the one based on L45 database will be named L45-MOD.

Errors (Figures 2 and 3) generated by L45-MOD and ATP-MOD are close but L45-MOD has a small advantage when applied – the MAE obtained by comparing the predicted coagulant dose and the actual one for test databases are lower. It is 5.47 g/m^3 (with 5.13 g/m^3 as SD) by applying the L45-MOD and 5.69 g/m^3 (with 5.47 g/m^3 as SD) by applying the ATP-MOD. Furthermore, applying ATP-MOD and L45-MOD, respectively, on their training database (Figure 3(a)) or applying ATP-MOD and L45-MOD separately on test databases (Figure 3(b) and 3(c)) show that

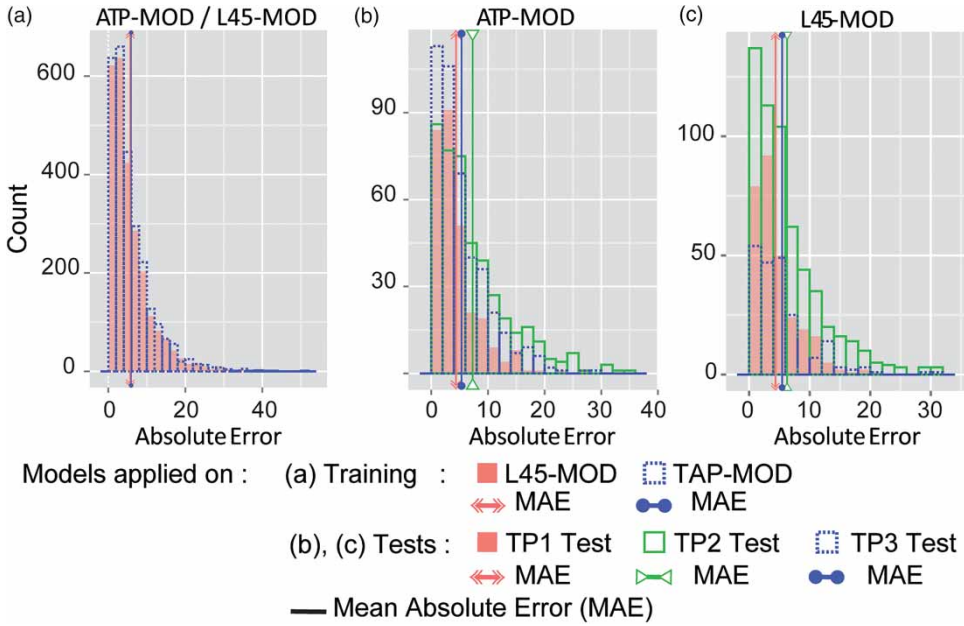


Figure 3 | Absolute error distribution and MAE of: (a) ATP-MOD and L45 MOD applied on the training database; (b) ATP-MOD applied on test databases; (c) L45-MOD applied on test databases.

more than 50% of all tested databases have an absolute error lower than 4.50 g/m^3. Moreover, it is clear in Figure 3 that the majority of the samples are concentrated around an absolute error close to zero.

Additionally, the coagulant doses computed with the models are very close to the actual data: the correlation coefficient between calculated and added actual doses reaches 0.85 for the ATP-MOD and 0.81 for the L45-MOD.

MAE generated by L45-MOD applied on test databases TP1, TP2, and TP3 are closer to each other than the ones generated by ATP-MOD (Figure 3, broken lines). Consequently, the L45-MOD generalizes well to new data (test data) for different types of WTP. Thus, L45-MOD can be used for predicting coagulant optimal dose when the raw water turbidity is lower than 45.5 NTU, otherwise ATP-MOD should be used.

For the treatment plant operators, these models are simple to use. They need only to input the real-time measurement of turbidity, conductivity, and pH of the raw water, and fix these same parameters for the treated water as needed, e.g., to meet local standards.

The use of L45-MOD or ATP-MOD is simple and provides a high accuracy of the results. The use of one or other model depends on raw water turbidity level: for raw water with low turbidity, not exceeding 45.5 NTU, the use of L45-MOD is recommended. However for sudden events (flood, rainstorm), during which the turbidity becomes higher than 45.5 NTU, it is advisable to use ATP-MOD.

ANN model development discussion

Modeling through ANN using large data improves the performances of the models but too large databases can have the opposite effects on the decrease of model errors (Ellis & Morgan 1999). Here, despite the large databases that exceed 900 samples for TP1 and 1,400 samples for TP2 and TP3, the developed models show low absolute errors but these rates could be lower because of the outliers that can have an impact on the MAE. Indeed, the outliers are exceptional cases that can generate high error values (absolute errors) influencing the average value (the MAE). However, the outliers do not seem to influence the results, as shown in Figure 3 and deduced from the SD values, since the obtained absolute errors are very close to the mean MAE.

The assessment of the models' performance can be done by comparing model error criteria to jar test doses minus real doses or the error criteria of other models. The

models developed here guarantee an MAE of 5.69 g/m^3 and reached a minimum of 4.35 g/m^3 for more than 700 test samples, values that are near or even lower than the jar test error; the doses of coagulant determined by jar test compared to the real doses added to the treatment process can differ by about 5 g/m^3 (Van Leeuwen *et al.* 1999; Maier *et al.* 2004). Moreover, ATP-MOD and L45-MOD generate an average SD of 5.13 g/m^3 and 4.87 g/m^3, respectively. The best SD achieved is 3.71 g/m^3, while Van Leeuwen *et al.* (1999) in his study selected a model to predict coagulant dose with a SD of 6.2 g/m^3 with the best SD at 4.7 g/m^3. The best SD reached by the model developed by Maier *et al.* (2004) was 3.9 g/m^3 (for only 32 points). The comparison of the built models' error criteria to jar test error or to the error criteria of other models show that ATP-MOD and L45-MOD have improved the performances of the existing techniques or models.

The models developed in this paper and those cited above have similarly used raw and treated water parameters as inputs, to predict the dose of coagulant. However, the data used by Van Leeuwen *et al.* (1999) and Maier *et al.* (2004) and those used in this study are different; the two researchers used fewer samples from rivers, 202 samples for Maier *et al.* (2004) and only 40 samples for Van Leeuwen *et al.* (1999), and they both compared the results predicted by the model to the jar test doses and not to real added doses. Unlike these two models, ATP-MOD and L45-MOD are developed according to a large training data based on real experience (real added doses of coagulant during the treatment process).

The models that are developed in this study have improved performance compared to models developed by Van Leeuwen *et al.* (1999), Maier *et al.* (2004) and better results in terms of errors than the developed models by Park *et al.* (2008) and Hernandez & Le Lann (2006) that have a RMSE of 20.84 for the first one and 24.90 for the second, while the calculated RMSE for ATP-MOD and L45-MOD is clearly lower and equal to 7.32.

Thus, ATP-MOD and L45-MOD can be adopted to predict optimal dose of coagulant to be used in different WTP, with a guarantee of low errors. However, Minns & Hall (1996) affirmed that ANNs are unable to be extrapolated beyond the range of the data used for training. Despite the use of large ranges and the proved generalization capacity of the developed models, the databases are confined to their values (e.g., turbidity does not exceed 318 NTU) and Minns & Hall's observations could open up other perspectives for our results by testing the adopted models on other treatment plants presenting very high raw water conductivity, pH, and turbidity.

CONCLUSION

Two artificial neural network models are developed to predict the optimal dose of coagulant to be added during the coagulation flocculation process of drinking water treatment with the particularity to use common parameters easily measured or fixed as inputs: turbidity, conductivity, and pH of raw and treated water. These inputs are selected among others based on PCA results that are confirmed by HCL. It was shown that the performance of the artificial neural network depends on the training database.

The developed models are distinguished by their ability to be applied at various treatment plants with different treatment processes, a deduction based on the calculated errors that are low and close to each other whatever the water treatment database tested. The L45-MOD model, specific to raw water with a turbidity lower than 45.5 NTU, predicted coagulant doses that are highly correlated to real doses (a correlation coefficient equal to 0.81), with a MAE of about 5.47 g/m^3 and a SD equal to 4.87 g/m^3. Another model would be used, ATP-MOD, for raw waters with a turbidity higher than 45.5 NTU. This model likewise shows predicted doses highly correlated to real ones (a correlation coefficient equal to 0.85) with a MAE of 5.69 g/m^3 and SD of 5.13 g/m^3.

Regionally, in Tunisia, a software using the models' equations will soon be implemented to be a useful tool for the operators of the studied WTP, replacing manual dosage and providing a real-time coagulant dose. It can also be a tool to assess and compare the consumption of chemicals with that of the previous year, especially knowing that reduction of excessive doses of coagulant can reduce soda consumption needed for pH adjustment and increase the filters' longevity. In other countries, the models can be used thanks to their capacity of generalization (low errors for many types of water treatment process), and adjusted

flexibly by fixing the treated water inputs (treated water turbidity, conductivity, and pH) according to local standards. This would help in time and cost savings as well as potable water quality improvement.

REFERENCES

Arlot, S. & Celisse, A. 2010 A survey of cross-validation procedures for model selection. *Statistics Surveys* **4**, 40–79.

Astel, A., Biziuk, M., Przyjazny, A. & Namieśnik, J. 2006 Chemometrics in monitoring spatial and temporal variations in drinking water quality. *Water Research* **40** (8), 1706–1716. doi:10.1016/j.watres.2006.02.018.

Baba, K., Enbutu, I. & Yoda, M. 1990 Explicit representation of knowledge acquired from plant historical data using neural network. In: *IJCNN International Joint Conference on Neural Networks*. IEEE, pp. 155–160. doi:10.1109/IJCNN.1990.137838.

Baouab, M. H. & Cherif, S. 2015 Changement climatique et ressources en eau: tendances, fluctuations et projections pour un cas d'étude de l'eau potable en Tunisie [Climate change and water resources: trends, fluctuations and projections for a case study of potable water in Tunisia]. *La Houille Blanche* **5**, 99–107.

Baouab, M. H. & Cherif, S. 2017a Revolution impact on drinking water consumption: real case of Tunisia. *Social Indicators Research* **132** (2), 841–859.

Baouab, M. H. & Cherif, S. 2017b Identification of indispensable components for a better drinking water quality management: Tunis case of study. *Journal of Hydroinformatics* **19** (6), 942–952. doi:10.2166/hydro.2017.070.

Bazer-Bachi, A., Puech-Coste, E., Ben Aim, R. & Probst, J. L. 1990 Modélisation mathématique du taux de coagulant dans une station de traitement d'eau [Mathematical modelling of optimal coagulant dose in water treatment plant]. *Revue des Sciences de l'Eau* **3** (4), 377–397. doi:10.7202/705081ar.

Dębska, B. & Guzowska-Świder, B. 2011 Application of artificial neural network in food classification. *Analytica Chimica Acta* **705** (1), 283–291. doi:10.1016/j.aca.2011.06.033.

Delgrange-Vincent, N., Cabassud, C., Cabassud, M., Durand-Bourlier, L. & Laine, J. M. 2000 Neural networks for long term prediction of fouling and backwash efficiency in ultrafiltration for drinking water production. *Desalination* **131** (1), 353–362. doi:10.1016/S0011-9164(00)90034-1.

Dreyfus, G. 2005 *Neural Networks: Methodology and Applications*. Springer Science & Business Media. Original French edition published by Eyrolles, Paris, 497 p.

Ellis, D. & Morgan, N. 1999 Size matters: an empirical study of neural network training for large vocabulary continuous speech recognition. In: *Proceedings, 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2. IEEE, pp. 1013–1016.

Faraway, J. & Chatfield, C. 1998 Time series forecasting with neural networks: a comparative study using the airline data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **47** (2), 231–250. doi:10.1111/1467-9876.00109.

Gagnon, C., Grandjean, B. & Thibault, J. 1997 Modelling of coagulant dosage in a water treatment plant. *Artificial Intelligence in Engineering* **11** (4), 401–404. doi:10.1016/S0954-1810(97)00010-1.

Guhl, F. & Brémond, B. 2000 Optimisation du fonctionnement des réseaux d'eau potable. Prise en compte de l'aspect stochastique de la demande [Optimization of drinking water networks operation. Considering the stochastic aspect of the demand]. *Ingénieries-EAT* **23**, 15–23.

Hair, J. F., Black, B., Anderson, R. & Tatham, R. 1995 *Multivariate Data Analysis: Text and Readings*, 7th edn. Pearson, London, 816 p.

Hecht-Nielsen, R. 1989 Theory of the backpropagation neural network. In: *IJCNN. International Joint Conference on Neural Networks*. IEEE, pp. 593–605.

Heddam, S., Bermad, A. & Dechemi, N. 2012 ANFIS-based modelling for coagulant dosage in drinking water treatment plant: a case study. *Environmental Monitoring and Assessment* **184** (4), 1953–1971. doi:10.1007/s10661-011-2091-x.

Hernandez, H. & Le Lann, M. V. 2006 Development of a neural sensor for on-line prediction of coagulant dosage in a potable water treatment plant in the way of its diagnosis. In: *Advances in Artificial Intelligence-IBERAMIA-SBIA*. Springer, Berlin, Heidelberg, pp. 249–257. doi:10.1007/11874850_29.

Hornik, K., Stinchcombe, M. & White, H. 1989 Multilayer feedforward networks are universal approximators. *Neural Networks* **2** (5), 359–366. doi:10.1016/0893-6080(89)90020-8.

Kabsch-Korbutowicz, M. & Kutylowska, M. 2011 Use of artificial intelligence in predicting the turbidity retention coefficient during ultrafiltration of water. *Environment Protection Engineering* **37** (2), 75–84.

Koffi, Y. B., Ahoussi, K., Kouassi, A., Christophe, K. L. & Biemi, J. 2012 Modélisation de la consommation en eau potable dans les capitales Africaines au Sud du Sahara: application des reseaux de neurones formels à la ville de Yamoussoukro, capitale politique de la Côte D'ivoire [Modeling of drinking water consumption in African capitals in south of the Sahara: application of formal neural networks to the city of Yamoussoukro, political capital of the Ivory Coast]. *Journal of Asian Scientific Research* **2** (10), 562–573.

Kumar, J. S., Poongodi, P. & Balakumaran, P. 2013 Artificial intelligence based alum dosage control in water treatment plant. *International Journal of Engineering and Technology* **5** (4), 3344–3350.

Lamrini, B., Benhammou, A., Karama, A. & Le Lann, M. V. 2005 A neural network system for modelling of coagulant dosage used in drinking water treatment. In: *Adaptive and Natural Computing Algorithms* (B. Ribeiro, R. F. Albrecht, A. Dobnikar, D. W. Pearson & N. C. Steele, eds). Springer, Vienna, pp. 96–99. doi:10.1007/3-211-27389-1_23.

Li, X. 1996 Simultaneous approximations of multivariate functions and their derivatives by neural networks with one hidden layer. *Neurocomputing* **12** (4), 327–343. doi:10.1016/0925-2312(95)00070-4.

Li, P., Lin, K., Fang, Z. & Wang, K. 2017 Enhanced nitrate removal by novel bimetallic Fe/Ni nanoparticles supported on biochar. *Journal of Cleaner Production* **151**, 21–33.

Liano, K. 1996 Robust error measure for supervised neural network learning with outliers. *IEEE Transactions on Neural Networks* **7** (1), 246–250.

Maier, H. R. & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software* **15** (1), 101–124. doi:10.1016/S1364-8152(99)00007-9.

Maier, H. R., Morgan, N. & Chow, C. W. 2004 Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. *Environmental Modelling & Software* **19** (5), 485–494. doi:10.1016/S1364-8152(03)00163-4.

Minns, A. W. & Hall, M. J. 1996 Artificial neural networks as rainfall-runoff models. *Hydrological Sciences Journal* **41** (3), 399–417. doi:10.1080/02626669609491511.

Park, S., Bae, H. & Kim, C. 2008 Decision model for coagulant dosage using genetic programming and multivariate statistical analysis for coagulation/flocculation at water treatment process. *Korean Journal of Chemical Engineering* **25** (6), 1372–1376. doi:10.1007/s11814-008-0225-9.

Rak, A. 2013 Water turbidity modelling during water treatment processes using artificial neural networks. *International Journal of Water Sciences* **2**, 3. doi:10.5772/56782.

Rakotomalala, R. 2008 *Tests de normalité: Techniques empiriques et tests statistiques [Normality Tests: Empirical Techniques and Statistical Tests].* Version 2.0. Université Lumière Lyon, Lyon, France. 59 p.

Rodriguez, M. J. & Serodes, J. B. 1996 Neural network-based modelling of the adequate chlorine dosage for drinking water disinfection. *Canadian Journal of Civil Engineering* **23** (3), 621–631. doi:10.1139/l96-871.

Shetty, G. R. & Chellam, S. 2003 Predicting membrane fouling during municipal drinking water nanofiltration using artificial neural networks. *Journal of Membrane Science* **217** (1), 69–86.

Sockloff, A. L. N. & Edney, J. 1972 *Some Extension of Student's t and Pearson's r central Distributions*. Technical Report. Measurement and Research Center, Temple University, Philadelphia, PA, USA.

Tabachnick, B. G. & Fidell, L. S. 2007 *Using Multivariate Statistics*. Pearson Education, Boston, MA, USA.

Takahama, S., Wittig, A. E., Vayenas, D. V., Davidson, C. I. & Pandis, S. N. 2004 Modeling the diurnal variation of nitrate during the Pittsburgh Air Quality Study. *Journal of Geophysical Research: Atmospheres* **109** (D16). doi:10.1029/2003JD004149.

Taylor, K. E. 2001 Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres* **106** (D7), 7183–7192. doi:10.1029/2000JD900719.

Valentin, N. 2000 *Construction d'un capteur logiciel pour le contrôle automatique du procédé de coagulation en traitement d'eau potable [Construction of a Software Sensor for the Automatic Control of the Coagulation Process in Drinking Water Treatment].* Doctoral thesis, UTC/L.desEaux/CNRS, France.

Van Leeuwen, J., Chow, C. W. K., Bursill, D. & Drikas, M. 1999 Empirical mathematical models and artificial neural networks for the determination of alum doses for treatment of southern Australian surface waters. *Aqua* **48**, 115–127. doi:10.1046/j.1365-2087.1999.00135.x.

Vítkovský, J. P., Simpson, A. R. & Lambert, M. F. 2000 Leak detection and calibration using transients and genetic algorithms. *Journal of Water Resources Planning and Management* **126** (4), 262–265. http://dx.doi.org/10.1061/(ASCE)0733-9496(2000)126:4(262).

Williams, B., Brown, T. & Onsman, A. 2012 Exploratory factor analysis: a five-step guide for novices. *Australasian Journal of Paramedicine* **8** (3), 1.

Wu, G. D. & Lo, S. L. 2008 Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. *Engineering Applications of Artificial Intelligence* **21** (8), 1189–1195. doi:10.1016/j.engappai.2008.03.015.

Zhao, S. & Xu, Y. 2005 Multivariate statistical process monitoring using robust nonlinear principal component analysis. *Tsinghua Science & Technology* **10** (5), 582–586. doi:10.1016/S1007-0214(05)70122-X.

First received 16 February 2018; accepted in revised form 11 July 2018. Available online 30 July 2018