Brazilian Journal
of Chemical
Engineering

ISSN 0104-6632
Printed in Brazil
www.abeq.org.br/bjche

Vol. 35, No. 04, pp. 1369 - 1381, October - December, 2018
dx.doi.org/10.1590/0104-6632.20180354s20170039

APPLICATION OF UNCERTAINTY ANALYSIS OF ARTIFICIAL NEURAL NETWORKSFOR PREDICTING COAGULANT AND ALKALIZER DOSAGES IN A WATER TREATMENT PROCESS

F. C. de Menezes[1], R. M. Fontes[1], K. P. Oliveira-Esquerre[1]* and R. Kalid[1]

[1] Program of Industrial Engineering (PEI),Department of Chemical Engineering, Polytechnic School,
Federal University of Bahia (UFBA), Rua Prof. Aristides Novis02, EP-UFBA, Federação, 40210-630, Salvador, BA, Brazil.

(Submitted: January 20, 2017 ; Revised: September 4, 2017 ; Accepted: October 23, 2017)

Abstract - Artificial neural networks (ANNs) were built to predict coagulant (Model I) and alkalizer (Model II) dosages given raw and treated water parameters from a water clarifying process. Different ANN architectures were tested and optimal results were obtained with [10-10-10-01] and [08-12-12-01] nodes of input, hidden and output layers for Models I and II, respectively. Two algorithms based on GUM-S1weredevelopedto evaluate the artificial neural network parameter uncertainty and the coverage interval of model outputs. The results show that these algorithms can provide a better set of parameters for the ANN compared with the traditional training method. The present research provides a unique unifying view that considers neural networks and uncertainty analysis in a well-documented industrial case study.
Keywords: Artificial intelligence; Parameter uncertainty; Coverage interval; Aluminum sulfate; Sodium hydroxide.

INTRODUCTION

For the past twenty years many authors have published results of coagulation process modeling aiming to predict the optimal coagulant dosage. This process is considered complex and not fully understood since the interactions between transfer mechanisms and chemical and biological kinetics happen in a poorly misunderstood environment (Maier et al., 2004; Dobias and Stechemesser, 2005);therefore, deterministic models are extremely difficult to develop (Wu and Lo, 2008).

Empirical models based on artificial neural networks (ANNs) have been developed to estimate hard-to-measure process variables.ANN has proven to be able to empirically describe the nonlinear relationships between water characteristics and optimal coagulant dosages in coagulation processes. Table 1 highlights some papers where coagulant dosage was successfully estimated by ANN or other empirical technique.

Although estimation of uncertainty in the identification parameter is important in order to obtain a more representative model, uncertainty analysis of ANN parameters is not common in the literature. For instance, none of the works listed in Table 1 considered the uncertainty associated with the data nor with the model parameters. The only papers of ANN uncertainty analysis found in the literature were Shrestha et al. (2009) and Srivastav (2007). The former proposed an alternative way to evaluate the measurement uncertainty for ANN by using sensitivity analysis and the Monte Carlo method. The second evaluates the uncertainty of an ANN-based hydrological model using the bootstrap method. Methods to evaluate the uncertainty of measurements are well described in the guides BIPM et al. (2008a), BIPM et al. (2008b) and

* Corresponding author: Tel.: +55 71 3283-9800, Fax.: +55 71 3283-9802 - Email address: karlaesquerre@ufba.br

Table 1. Papers related to the prediction of coagulant dosage by using artificial neural networks.

Authors (year)	Modeling techniques	Data/information required
Heddam and Dechem (2015).	Neural-fuzzy inference system (DENFIS)	Seven online measurements of raw water and alum dosage
Heddan et al (2011)	Generalized regression neural network (GRNN) and radial basis neural network (RBFNN)	Six input variables 725 samples in 2 year-period
Heddam et al. (2012)	Adaptative neuro-fuzzy inference system (ANFIS)	Seven online measurements of raw water and alum dosage
Hernandez and Le Lann (2006)	ANN and linear regression model	9 water quality variables
Lamrini et al. (2005)	Neural-fuzzy inference system	Seven online measurements of raw water and alum dosage.
Robenson (2009)	Inverse neural network	
Wu and Lo (2008)	Levenberg-Marquardt neural networks	Raw water turbidity and coagulant dosage on day $t-1$
Zhang et al. (2013)	K-nearest neighbors (KNN)	3 water quality variables 966 samples collected between 2008 and 2010

BIPM et al. (2011). Furthermore, Lira (2011) proposes a method to evaluate the uncertainty in the estimation of the parameters of a model.

In this paper, ANN inverse models were developed for aluminum sulfate dosage (coagulant) prediction based on raw and treated water parameters (Model I).Then, inverse models were developed for sodium hydroxide (alkalizer) dosage prediction given the coagulant dosage; raw and treated water parameters (Model II). The relation of both models is based on the importance of pH adjustments when sodium hydroxide is applied to obtain the optimal dosage for coagulation of surface water with wide quality variation. The evaluation of the ANN parameter uncertainty, prediction of the coverage interval and estimate of the ANN output were carried out by uncertainty analysis based on the GUM-S2 approach (BIPM et al., 2011), which consists of the uncertainty propagation in multivariate systems based on the concept of covariance matrix and joint probability distribution function, extending the rules presented in GUM (BIPM et al. 2008a)and GUM-S1 (BIPM et al. 2008b).Models I and II with uncertainty analysis may be used together to represent the coagulation process of a water treatment plant. The present research provides a unique unifying view that considers neural networks and uncertainty analysis in a well-documented industrial case study.

WATER CLARIFYING PROCESS INSIGHTS

The clarifying process is an important step in water treatment. This process consists of coagulation, flocculation, sedimentation and filtration with the purpose of removing impurities such as suspended solids, colloidal material and microorganisms (including pathogens). Aluminum sulfate (the hydrated form is called alum: $Al_2(SO_4)_3.nH_2O$) is commonly used as a chemical agent for coagulation responsible for particle destabilization and aggregation.

pH is the main variable affecting the coagulation and flocculation performance of the water treatment process. The effect of the pH can be considered critical for coagulation due to the formation of hydrogen ions and hydrolyzed products which depend on the pH level (Yan et al. 2008) and better coagulation-flocculation can be achieved if the pH level is adequate. Concerning aluminum sulfate, in a lower pH the charge neutralization and adsorption mechanisms are favored (allowing better color removal), whereas in a higher pH the enmeshment mechanism is favored (promoting better turbidity removal) (Di Bernardo and Sabogal Paz, 2008). In addition, inadequate adjustment of pH can produce aluminum residues (Dorea, 2009), which may represent a health risk associated with Alzheimer's disease (Flaten, 2001).

The optimum pH remains almost constant, but the pH range becomes more restrictive as the coagulant dosage decreases. In other words, an optimal coagulant dosage can be obtained through optimal pH adjusting.

The coagulant dosage and pH adjusting depends on the quality of the water to be treated and can be obtained by jar testing. However, jar testing is relatively expensive and time-consuming (Joo et al., 2000; Maier et al., 2004; Wu and Lo, 2008), and consequently it can just be carried out periodically (Yu et al., 2000). Given this, water treatment plants are unable to respond quickly to changes in raw water quality (Joo et al., 2000) and it cannot be used for real-time control (Wu and Lo, 2008; Yu et al., 2000). With regard to adjusting the pH, the operators infer the ideal range of pH to operate the clarifying process, but significant changes in the raw water quality also can hinder the pH control system. Hence, regulating the pH becomes a difficult task without jar testing, as well as for prediction models that do not take into consideration pH control.

Some water treatment plants that are supplied by surface water face serious problems of water quality variation. For example, during a heavy rain

storm water carries sand, silt and organic particles/ compounds that increase suspended solids, color and turbidity, changing the raw water quality. Therefore, changes in surface water quality are relevant because water becomes more susceptible to pollution and contamination (Ouyang et al., 2006), requiring more robust plants (Di Bernardo and Sabogal Paz, 2008).

CASE STUDY

Historical data of jar test results were obtained from a water treatment plant located in the Industrial Complex of Camaçari, the largest industrial complex in Latin American. This complex covers approximately 5.700m^3/h and produces more than 11.5 million tons of primary, intermediate and final chemical and petrochemical products per year. The water treatment plant is responsible for producing clarified, filtered, demineralized water as well as drinking water. The Joanes River provides the water supply to this plant (Oliveira-Esquerre et al., 2009).

Jar tests are carried out periodically in the water treatment plant. First, different dosages are added to jars with samples of raw water, adjusting the amount of treatment chemicals (coagulant and alkalizer). Second, the samples are stirred until flocks are formed, developed and settled. The operator then performs a series of tests (in total eight) and observes the effects of the different dosages applied.

The dosage of chemical agents (a coagulant and a pH adjuster), quality parameters of raw water and treated water measures from jar tests carried out over a six-year period were used to develop and validate ANN models (enabling seasonal and operational patterns of this period to be captured by ANNs).Raw and treated water parameters were chosen based on their expected relation with coagulant and alkalizer dosages, as reported in the literature (see papers in Table 1). Some basic statistics of the variables of 1940sample-size are shown in Table 2.

The raw water quality parameters were: pH (pH_r), color (col_r), turbidity (turb_r), suspended solids (ss_r) and alkalinity (alka_r). The dosages of chemical agents are coagulant (aluminum sulphate - Sulf_in) and alkalizer (sodium hydroxide - NaOH_in). The quality parameters of the treated water were pH (pH_t), color (col_t), turbidity (turb_t), suspended solids (ss_t) and residual aluminium (alum_t), in and t are related to raw, influent and treated water. The Anderson-Darling (AD) normality test was performed for a 90% confidence interval. All variables exceeded the critical value of the p-value (equal to 0.10), so the hypothesis of normality was rejected. Skewness and kurtosis were also evaluated to check the normality of the data.

The available data are shown in Figure 1.Color and turbidity show seasonal behavior. Peaks are related to the occurrence of heavy rain that promotes changes in raw water quality and consequently makes it difficult to control the treated water quality. In such cases, Zhang and Stanley (1997) have shown water quality is much more difficult to predict by ANN. The observed pH (pH_r) decrease is related to industrial effluents released into the water upstream. The alkalinity (alka_r) has widened its range of variation in recent years. Color (color_r) varies widely as can be observed in the time series and basic statistics, and efficient color removal depends on the pair of optimum pH and coagulant dosage. These data behavior were not present in other studies which focused on the modeling of coagulant dosage through quality parameters of raw water (Joo et al, 2000; Maier et al., 2004).

ANN MODELING

Because neural networks are massively parallel, they have a better filtering capacity and generally perform better than traditional linear models with noisy or incomplete data (Baughman, 1995). The trained model may be continuously adapted to new data, even though neural networks require previous data to be useful. In this research, neural network models were built and validated to predict the alkalizer dosage and coagulant dosages based on their relations with water

Table 2. Basic statistics for the variables available.

	Parameter	Minimum	Maximum	Average	S.D.	Skewness	Kurtosis	Unit
Raw water	pH_r	5.7	8.5	6.5	0.4	1	3.3	-
	col_r	11	696	132	86	2.2	8.3	(HU)
	ss_r	2.0	84,0	9,3	7,9	4.9	36.4	(mg/L)
	turb_r	0.5	38,0	11,5	7,3	1.3	1.9	(NTU)
	alka_r	8	24	13	3	0.9	0.6	(mg/L)
Dosage of chemical agents	NaOH_in	0.0	35,0	9,3	6,3	0.9	0.5	(mg/L)
	Sulf_in	5.0	75,0	33,9	10,9	0.6	-0.7	(mg/L)
Treatedwater	pH_t	4.5	7,2	6,1	0,3	-0.7	3.8	-
	col_t	1	102	8	6	6.3	73.7	(HU)
	ss_t	1.0	12.0	1.5	1.1	3.6	17.7	(mg/L)
	turb_t	0.1	9,1	0,8	0,6	5.5	56.6	(NTU)
	alum_t	0.000	1.000	0.091	0.002	6.1	50.8	(mg/L)

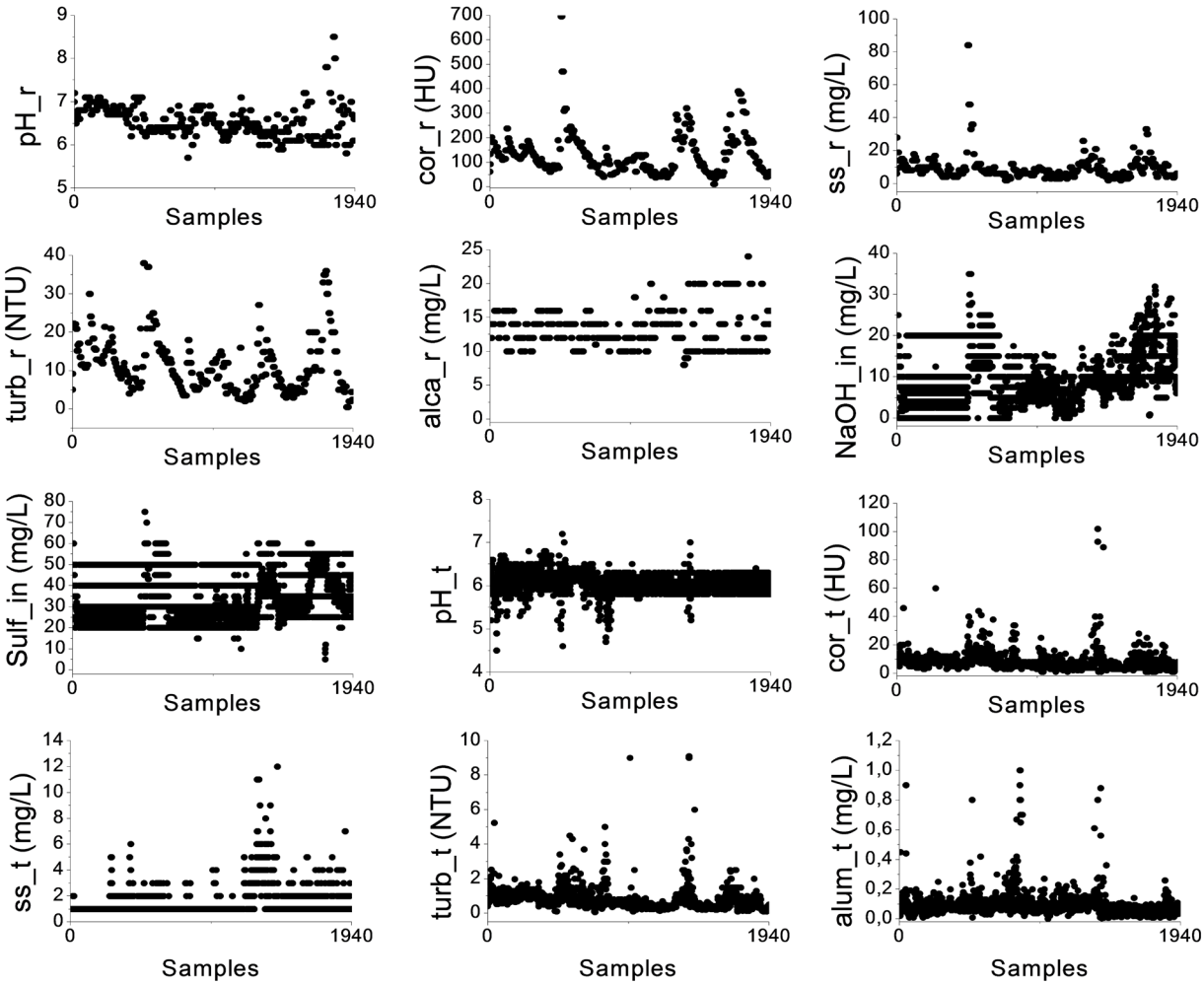


Figure 1. Plot of the available data over a six-year period.

quality parameters. Table 3 shows the description of input and output variables used to predict alkalizer dosage (NaOH_in) and coagulant dosage (Sulf_in).

Data Pre-Processing

Environmental data commonly contain samples that are different from the collection data and can be identified by some methods, for example: time series analysis, quartile method and Principal Component Analysis (PCA)(De Bruyne et al, 2006). The first enables observing data that do not follow the general trends and seasonality (Johnson and Wichern, 2007). The second is a method based on the position of a sample in relation to the upper and lower data limits of the one variable (Daszykowski et al., 2007). The third, PCA, is a multivariate method widely used for outlier analysis and variable orthogonalization (Chen et al., 2009; Daszykowski et al., 2007), and has been successfully used to deal with environmental variable data (Singh, 1996).

PCA transformation is defined by a set of p-dimensional vectors of weights or loadings (w_k w, Equation 1) that map each row vector, $X_{(i)}$ of X to a new vector of principal componentscores (t_i t, Equation 2), given by Eq. 3, in such a way that individual variables that are considered over the data set successively inherit the maximum possible variance from x, with each loading vector w constrained to be a unit vector.

$$w_k = \left(w_1, ..., w_p\right)_k \tag{1}$$

$$t_i = \left(w_1, ..., w_p\right)_i \tag{2}$$

Table 3. Inputs and output of the ANN inverse models.

Model		Inputs	Output
(I)	Process inputs	pH_r, col_r, turb_r, ss_r, alka_r, NaOH_in	Sulf_in
	Process outputs	col_t, turb_t, ss_t, alum_t	
(II)	Process inputs	pH_r, col_r, turb_r, ss_r, alka_r, Sulf_in	NaOH_in
	Process outputs	col_t, alum_t	

$$t_{k(i)} = x_i \cdot w_k \quad (3)$$

ANN Architecture

ANN architecture is composed of a multilayer perceptron (MLP), which has already been successfully used for the prediction of coagulant dosage (Maier *et al.*, 2004; Robenson *et al*., 2009; Wu and Lo, 2008).The number of neurons in an ANN should increase according to the complexity of the problem; nevertheless, a higher number of parameters (neurons) in relation to the number of available data for model training may compromise the ability of generalization. On the other hand, if there are insufficient neurons, it is difficult to obtain convergence during training (Maier and Dandy, 2000).

Feng et al. (2005) pointed out that the number of hidden layers depends on the complexity level of the relationship between inputs and outputs. Networks with one and two hidden layers were tested and the number of neurons in the hidden layer and the network geometry were found by trial and error. Two hidden layers may be required when the model with one layer does not show a good performance to represent the nonlinear input-output relations.

The optimization of the ANN is obtained by estimating the connected weights in a process called training. The Levenberg-Marquardt algorithm was used for training, which is one of the backpropagation algorithm modifications. Basically, the Levenberg–Marquardt algorithm performs a combined training process: around the area with complex curvature, the algorithm switches to the steepest descent algorithm, until the local curvature is proper to make a quadratic approximation; then it approximately becomes the Gauss–Newton algorithm, which can speed up the convergence significantly (Yu and Wilamowski, 2011).

The sigmoid transfer function is the most commonly adopted function for the hidden and output nodes (Lingireddy and Brion, 2005; Maier and Dandy, 2000). So, this function was used in ANN hidden layers, while a linear function was used in the output layer. Input data was normalized from 0.1 to 0.9 in order to avoid both the flat-spots of the sigmoidal function near its 0.0 and 1.0 limits and the loss of non-linear information when using a shorter interval, for example 0.2 to 0.8 (Oliveira-Esquerre et al.,2004).

Data Division

The main, well-known disadvantages of neural network training are that it requires large quantities of experimental data and the training of the network can take too long to be practical. Furthermore, its nonlinear approximation function can cause local minimum problems.

After pre-processing, the available data were divided into three sets: training, validation and testing. The training and validation sets were used to develop the models. The first one for adjusting the connection weights and the second to determine when to stop the training and to avoid overfitting (Oliveira-Esquerre et al., 2004).The test set was reserved to test the generalization ability of the model with data that were not used in the development process of the models.

In order to guarantee that the three sets contain the same statistical population and data from all the sampling period, data were sorted according to the date of the jar test. After randomization, 60%, 25% and 15%of each data set were selected for the training, validation and test set – respectively. This ratio was adopted based on the procedure of division suggested by Baxter et al. (2002), although some research indicate the use of up to 80% of samples for trainning (Lingireddy and Brion, 2005).

The data of the test and validation sets extrapolated to the training data limits were migrated to this data set, taking into account that ANN, as an empirical method, is not usually able to extrapolate (Oliveira-Esquerre et al., 2009). The maximum, minimum, interquartile range of the data and scores of principal components were evaluated to check if the procedure of data division was satisfactory to ensure that the three sets had the same statistical population. The training, validation and testing data sets, composed of 1166, 467 and 277, respectively, were then obtained.

The relative performance of the models was assessed through the coefficient of multiple determination (R^2) and the adjusted coefficient of multiple determination ($R^2_{adj.}$). The latter provides the coefficient of multiple determination taking into account the number of connection weights which compose the ANN model, and the degree of freedom (Oliveira-Esquerre et al., 2004).Furthermore, two measures of absolute error were used: mean absolute error (MAE) and root mean square error (RMSE).Both have been used to assess model performance when dealing with hydrological and water quality variables (Hamel and Smith, 2007), although some literature states that MAE is a more natural and unambiguous measure of average error (Willmott and Matsura, 2005).

UNCERTAINTY EVALUATION

In addition to the identification of the ANN model, a method to evaluate the uncertainty of the ANN parameters was developed. The method consists of applying the multivariate law of propagation of the probability density function (MLPP) approach to obtain the joint probability density function (PDF) of the parameters (BIPM et al, 2011).

The MLPP approach is recommended for the uncertainty evaluation of the measurand, which is represented by a non-linear measurement function. By using the joint PDF of the input, it is possible to evaluate the joint PDF of the measurand using the MLPP, as represented by Equation 4.

$$g_Y(\eta) = \int \cdots \int g_X(\xi) \cdot \delta\left[h(\eta;\xi)\right] d\xi_N \ldots d\xi_1 \quad (4)$$

where δ(.) denotes the Dirac function; ξ and η are the possible values of the input (X) and output (Y) variables; $g_Y(\eta)$ and $g_X(\xi)$ are the joint PDFs of the input (X) and output (Y) variables; h(.) is the implicit multivariate measurement function(BIPM et al, 2011).

This approach may be considered as being more complete, because it considers all the nonlinearity of the model and all the possible information of each variable. However, in some cases, the integrals are difficult to solve analytically, which requires a numeric method, such as the Monte Carlo Method (MCM) (BIPM et al, 2008; BIPM et al, 2011).

The MCM algorithm applied to solve the MLPP approach is summarized in the following steps:

1. Define the joint PDF, $g_X(\xi)$, for the input variables (X);
2. Define the number of Monte Carlos trials (M);
3. From the joint PDF, $g_X(\xi)$, M vectors (x_1, ..., x_M) are sampled;
4. Evaluate the non-linear measurement function, ($h(Y; X) = 0$) for each sample (x_1, ..., x_M) to obtain a M vector of the measurand (y_1, ..., y_M).
5. Compute the estimate and the covariance matrix from the measurand vector (y_1, ..., y_M).

Uncertainty Evaluation of the Parameters

The estimation of the parameters is an optimization problem where the decision variables are the parameters of the model and the cost function, in most cases, is the Least Square (LS). This cost function is a simplification of the Maximum Likelihood Estimation (MLE), as the weighted Least Square (WLS). The main difference between LS and WLS is that the former assumes that the experimental variance is constant, while in the latter this condition is not assumed.

The following conditions are required to carry out our uncertainty evaluation of the parameters: the variance of the output variable is constant and is known; variability of the input is much smaller than the variability of the output; the measurements of the inputs do not influence the measurement of the outputs; and all measurements are independent (Schwaab and Pinto, 2007).

In most commercial solvers to train an ANN, some cost function similar to the LS is used, like the MSE (Mean Square Error) or the determination coefficient (R^2). However, in most cases, the ANN is used to represent some systems, in which the list of conditions is not tested or is not obeyed, which may provide a result with little statistical meaning. A simplified algorithm is represented in Figure 2.

From Table 2, it is possible to make a PDF to fit to outputs (Sulf_in, NaOH_in) and the input variables; however, this is not the objective of this paper. It is also possible to know the Type A uncertainty evaluation that represents the variability (standard deviation) for each variable. The Type B uncertainty was not evaluated because there was not enough reliable information, such as calibration certificates, temperature influence on measurement systems, etc. Therefore, it was assumed that each variable is represented by a uniform PDF, with variance represented by Equation 5.

$$u_c(X_i) = \frac{x_i^{\max} - x_i^{\min}}{\sqrt{12}} \quad (5)$$

where $u_c(X_i)$ is the combined uncertainty of any variable X_i; $X_i^{\max}$ and $X_i^{\min}$ are, respectively, the maximum and minimum of any variable.

After this, the Monte Carlo Simulation was performed to solve the MLPP. This is a simple algorithm, which consists of evaluating the measurement function M (Monte Carlo sample) times.

In the training of the ANN, the measurement function is the optimization problem, and the measurand is the parameters of the ANN, so it is necessary to get M training samples from output variables and to solve the optimization problem for each one. In the end, M samples of the joint PDF of the parameters are obtained, with which it is possible to evaluate statistical moments, such as the mean, standard deviation, and others.

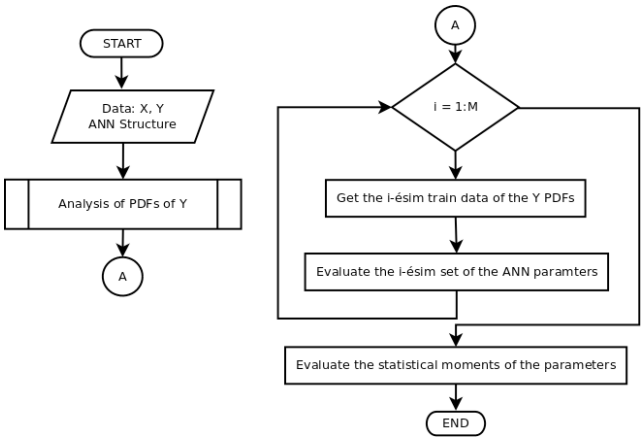


Figure 2. Simplified algorithm for the uncertainty evaluation of the parameters.

Propagation of the Uncertainty

From the propagation of the uncertainty it is important that the measurand (Y) is represented by a

perfect measurement function. Otherwise, measurand uncertainty will depend on the uncertainty parameters of the measurement function as well.

After the uncertainty evaluation of the parameters, this is propagated with the input variable uncertainties to find the output uncertainty. In this case, the Monte Carlo Method is also used; however, the measurement function is the ANN with the uncertainty in the parameters and input variables. Figure 3 shows the simplified MCM algorithm applied to obtain the measurand PDF.

To obtain the measurand PDF, it is necessary to get M validation samples of the input variables and the M samples of the parameters from their PDF. The ANN is simulated from each sample and, in the end, the measurand PDF is constructed, making it possible to calculate the statistical moments.

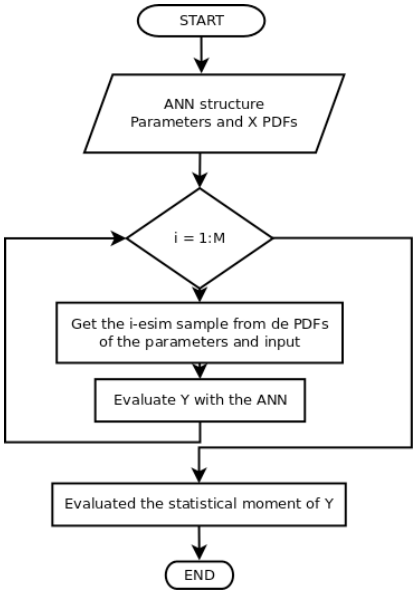


Figure 3. Simplified MCM algorithm applied to uncertainty propagation.

RESULTS AND DISCUSSION

Descriptive Analysis of the Data

Table 4 shows the loadings of the principal components and their eigenvalues. It shows that the outliers identified by using the first two methods were confirmed with PCA. Through scatter plotting of the PCA scores of the first and second principal components, it is possible to verify that there are two groups of observations (A and B)away from the mass of data (Figure 4).

Group A represents those days on which observations showed the highest values of raw water quality parameters. The measure of color and suspended solids on these days were about twice as high as the second highest measure. Group B represents the

Table 4. Loadings and eigenvalues of the principal components.

Principal component	Variance (%)	Cumulative variance (%)
1	34.7	34.7
2	19.4	54.1

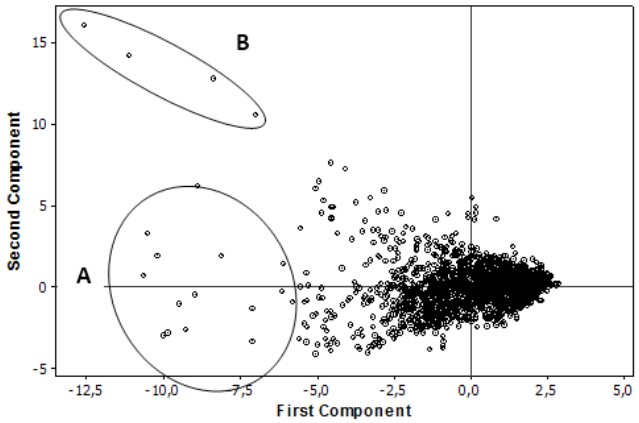


Figure 4. Result of PCA showing the scores of the first and second principal components.

results of jar testing that show high values of treated water quality parameters. Both groups were excluded. Since these conditions rarely happen during process operation, the data were reduced to 1910 samples at the end of the outlier analysis.

A correlation analysis of the model inputs was carried out considering a multivariate and bivariate analysis. For the multivariate analysis, a graph of the loadings of the first and second principal component is shown in Figure 5. It can be seen that there is an agreement of the physical relationship between color, suspended solids and turbidity, as well as between pH and alkalinity. The correlation of the application of alkalizer (NaOH_in) as a function of the coagulant (Sulf_in) is also verified in Figure 5.

For the bivariate analysis, the Spearman rank coefficient was used to assess the degree of correlation, linear or not, without requiring normalization of the

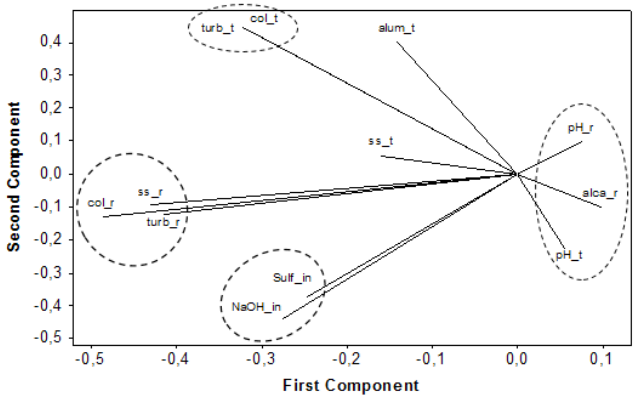


Figure 5. Graph of loadings corresponding to the first and second principal component.

data (Hogg and Craig, 1995; Montgomery and Runger, 2015), see Table5.The results agree with those obtained by using PCA, that is, the highest correlations were found for the same variables clustered in Figure 5.In addition, pH_r, col_r, turb_r, ss_r, alka_r, and ss_tshow significant correlation with Sulf_in and NaOH_with 95% confidence interval. In this case, a 95% confidence interval was used so that the identification phase of the potential variables was not heavily penalized. Sulf_in was included in the models due to its importance in drinking water. Despite the low correlation between NaOH_in and col_t, col_t was considered as an input of the alkalizer prediction model for the importance of monitoring and control of this parameter.

Modeling Results Before Uncertainty Evaluation

The best network topologies found for prediction of aluminum sulfate dosage (Model I) and sodium hydroxide dosage (Model II) are composed of[10-10-10-01] and [08-12-12-01] nodes in the input, hidden and output layers, respectively. Scatter plots of the measured versus predicted values of aluminum sulfate dosage (Sulf_in) and sodium hydroxide dosage (NaOH_in) for the testdata set are shown in Figures6 (a) and (b) – for a prediction interval of 90%. The prediction performance indices (R^2, R^2_{adj}MAE and RMSE) for the ANN models are shown in Table 6.The values show good agreement between predictions and measured data.

The plots of predicted and measured aluminum sulfate and sodium hydroxide are shown in Figures 6 (a) and (b). Both models provide predictive capability and, as expected, predict most outputs within 90% prediction bands.

Comparing the obtained results with others, Maier et al. (2004) found aR^2 of 0.94 and MAE of 3.2 mg/L. The results obtained in this research show R^2 smaller than 0.94 probably because of the bigger range of the raw water quality parameters, which indicates that the present research shows good results. Although both used jar test results, the seasonal variations in raw water and the pH adjustment were not taken into account in the cited study. Wu and Lo (2008) obtained a R^2 around 0.75 when taking into account seasonal variation.

Table 6. Indicators of model performances.

Model	R^2	R^2_{adj}	MAE	RMSE
I	0.77	0.77	3.79	5.17
II	0.81	0.81	1.99	2.68

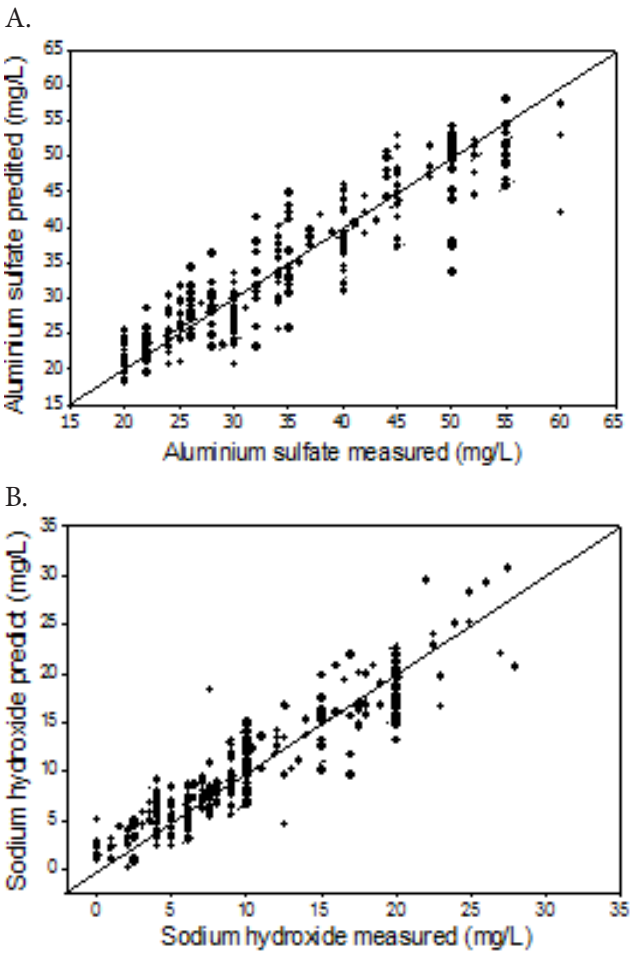


Figure 6. Relationship between predicted vs. measured value employing Model I (a) and Model II (b) for the test set. The dashed lines indicate the 90% prediction interval.

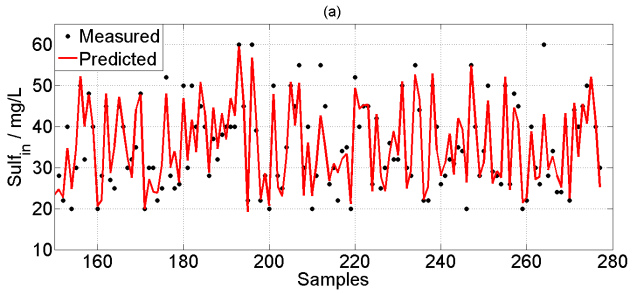
Table 5. Matrix of Spearman rank correlation coefficients.

	pH_r	col_r	ss_r	turb_r	alka_r	NaOH_in	Sulf_in	pH_t	cor_t	ss_t	turb_t
col_r	**-0.129**										
ss_r	-0.018	**0.772**									
turb_r	0.032	**0.880**	**0.732**								
alka_r	**0.051**	**-0.155**	**-0.066**	**-0.153**							
NaOH_in	**-0.269**	**0.341**	**0.225**	**0.259**	**-0.12**						
Sulf_in	**-0.231**	**0.255**	**0.174**	**0.156**	**-0.056**	**0.837**					
pH_t	**0.129**	**0.047**	-0.038	**0.066**	**0.046**	**0.114**	-0.038				
cor_t	**0.108**	**0.451**	**0.356**	**0.435**	**-0.11**	0.014	**-0.054**	**0.065**			
ss_t	**-0.16**	**0.164**	**0.213**	**0.053**	-0.036	**0.113**	**0.156**	-0.006	**0.119**		
turb_t	**0.164**	**0.511**	**0.435**	**0.502**	**-0.13**	-0.005	**-0.059**	**0.072**	**0.815**	**0.129**	
alum_t	**0.075**	-0.027	**0.101**	-0.019	-0.04	**-0.156**	-0.045	**-0.134**	**0.287**	**0.170**	**0.284**

In bold, coefficients of correlation among variables with statistical significance for a 95% confidence interval, i.e., *p-value*< 0.05.

To aid the evaluation of the model's predictive capacity, the predicted and measured values for the test data set were plotted for Model I and II, Figure 7(a) and (b), respectively. Both models show good results of forecasting, even for adverse operating conditions (i.e., with significant variations in the dose required). One must be aware that data interpolation should be avoided since some patterns were deleted during data pre-processing.

A.



B.

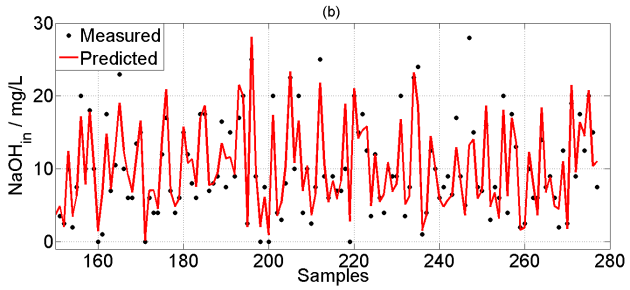


Figure 7. Plot of the predicted (point) and measured (solid line) values of aluminum sulfate dosage (a) and sodium hydroxide dosage (b) for samples 150 to 277 of the test set.

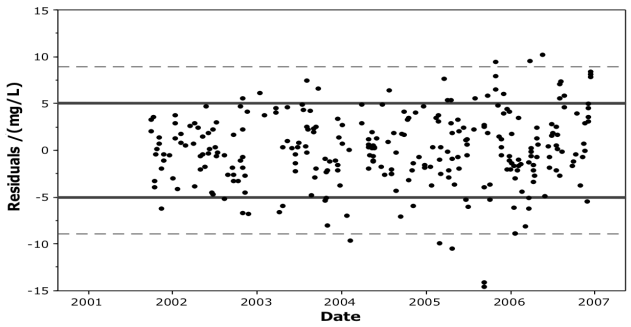
Times series plots of the model residuals are shown in Figure 8. No systematic pattern is observed. Model I shows about 83% residuals (for the test data set) within the variation limits adopted in the operation of coagulant dosage application (operation interval of 5.0 mg/L) in this water treatment plant –solid line. If this operation interval is compared with values of average model-performance error (MAE and RMSE), the good predictive capacity of the model is more evident. The dashed line indicates a 90% confidence interval. The residuals with the highest dispersion (since mid-2005) can be related to a decreasing pH in recent years, as well as widening of the alkalinity variation range.

Results with Uncertainty Evaluation

The ANN structure evaluated in the previous section, including the number of neurons and layers, was maintained to evaluate the parameter uncertainties.

Figure 9 shows the behavior of the output prediction with variations in the uncertainties for Model I (Sulf_in). As the uncertainty was not evaluated correctly,

A.



B.

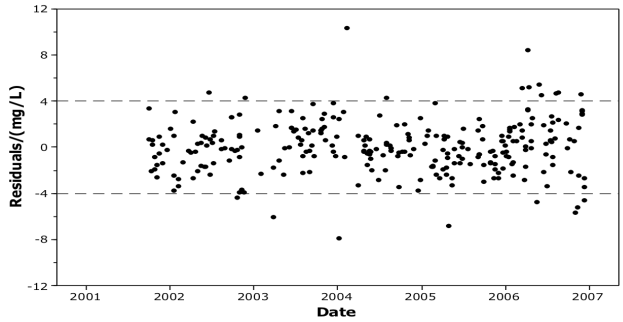


Figure 8. Residual time series of the Models I (a) and II (b) for the test set. Upper and lower dashed lines indicates 90% confidence interval and solid lines in (a) indicate variation limits adopted for the operation.

a sensitivity analysis using from half to double the uncertainty calculated by Equation (5) was used.

The output estimate (black line) and the confidence interval (95.45 %), represented by the minimum and the maximum (grays dash lines), are evaluated by using the PDFs predicted by the MCM method for each experimental sample (dots).

It is possible to observe that the increase in the output uncertainty implies in a decrease in the model performance, considering the mean of predicted PDFs. However, it should be noted that the predicted coverage intervals contain the experimental data, i.e., the evaluated PDFs consider the variability of the process.

The MCM method also allows us to find the joint PDF of the parameters and, with this, it is possible to evaluate any statistical moments. An ANN with the mean of the joint PDF of the parameters was simulated and the results are shown in Figure 10 and Table 7.

The results in Table 7, when compared with Table 6, show that the mean of the parameters provides little improvement in the results relative to traditional ANN training for Model I. Because "M" trainings were performed (in this paper 10^4 Monte Carlo samples were used), the mean of the parameters represents the estimate of the expected value of the parameters, which provides, in this case, a robust model.

For Model II, the same procedure was carried out and the results are shown in Figures 11 and 12 and Table 8.

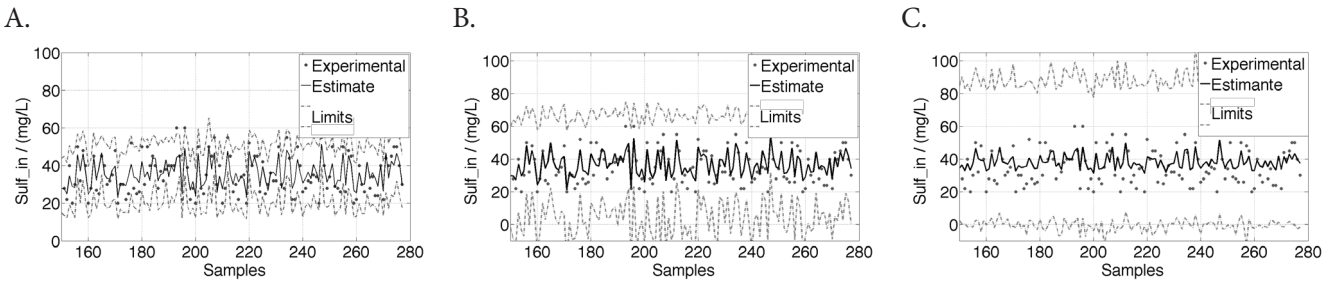


Figure 9. Aluminum sulfate estimated behavior with half (a), one time (b) and the double (c) of the uncertainty evaluated by Equation (5)**.**

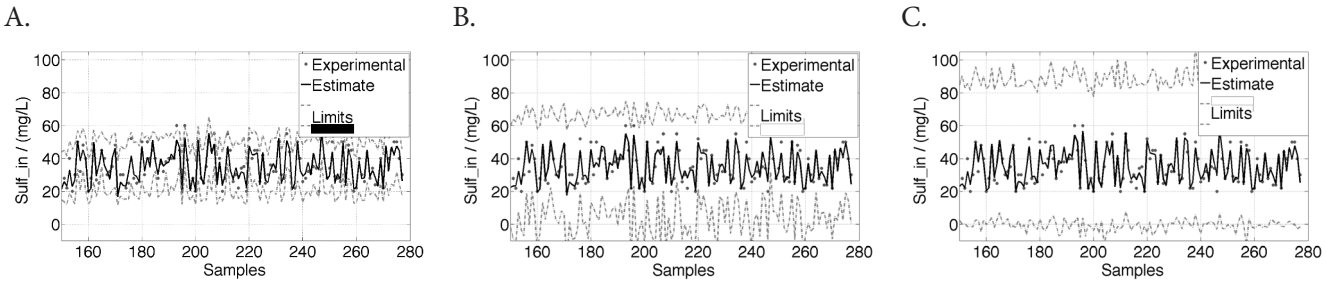


Figure 10. Aluminum sulfate prediction behavior with half (a), one times (b) and the double (c) of the uncertainty evaluated by Equation (5) and using the mean values of the parameters.

Table 7. Prediction performance indices by using the mean values of the parameters for Model I - MCM.

Uncertainty	R^2	R^2_{adj}	MAE / (mg/L)	RMSE / (mg/L)
$0.5u_c$	0.78	0.77	3.73	5.10
u_c	0.78	0.77	3.69	5.10
$2u_c$	0.76	0.75	3.82	5.33

Table 8. Values of the performance index with the mean of the parameters for Model II.

Uncertainty	R^2	R^2_{adj}	MAE / (mg/L)	RMSE / (mg/L)
$0.5u_c$	0.82	0.82	1.98	2.66
u_c	0.85	0.85	1.82	2.42
$2u_c$	0.87	0.87	1.82	2.42

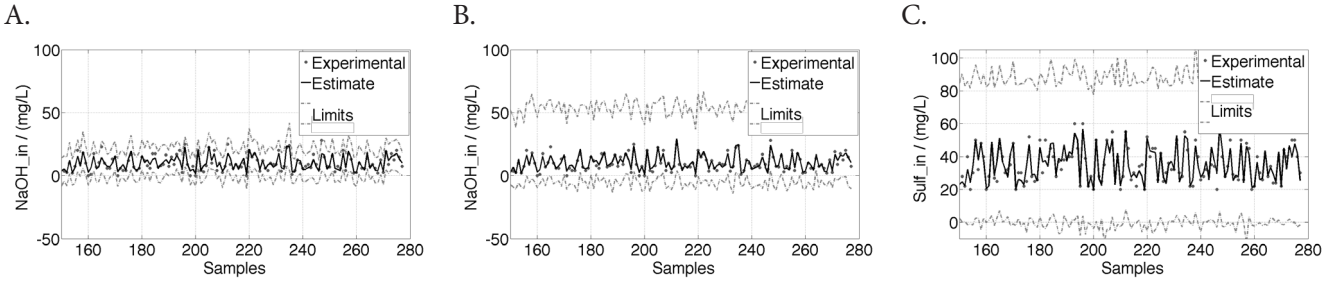


Figure 11. Sodium hydroxide estimated behavior with half (a), one times (b) and the double (c) of the uncertainty evaluated by Equation (5).

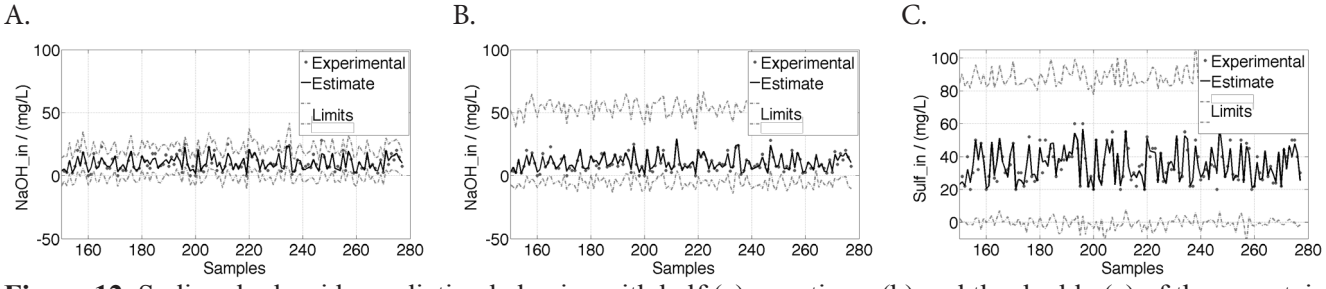


Figure 12. Sodium hydroxide prediction behavior with half (a), one times (b) and the double (c) of the uncertainty evaluated by Equation (5) and using the mean values of the parameters.

Both Model I and II performances are better than when uncertainty analysis is carried out, which indicates that this method can be used to improve the ANN performance.

CONCLUSIONS

ANN models for the prediction in real–time of coagulant (aluminum sulfate) and alkalizer (sodium hydroxide) dosage were developed for a clarifying process of a water treatment plant in Camaçari, Bahia, Brazil. The models took into account the quality parameters of raw and treated water, together with the process parameters and the wide quality variation of raw water. In this case, the quality that the water source has is associated with water pollution.

The ANN models without uncertainty analysis reproduce the aluminum sulfate and sodium hydroxide dosage based on jar test results satisfactorily. In the models developed, the prediction of the aluminum sulfate dosage depends on the sodium hydroxide dosage, and vice versa, which does not represent a difficulty for operators who have experience and empirical knowledge of the application of one in function of the other. Indeed, this already is done when the jar test is carried out. Model I considers alkalizer dosage variation and raw and treated water parameter measures to estimate the required dosage of coagulant. This model shows R^2, R^2_{adj}, MAE and RMSR equal to 0.77, 0.77, 3.79 mg/L and 5.17 mg/L, respectively. Slightly better results were obtained for Model II, withR^2, $R^2_{adj.}$, MAE and RMSR equal to0.81, 0.81, 1.99 mg/L; 2.68 mg/L, respectively. However, the coagulant is considered to be an input and the alkalizer an output of the model.

By applying the proposed algorithm to evaluate the parameters and the output uncertainty of the ANN, it was possible to verify that both models are sensitive to uncertainty. The results show that it is possible to obtain an output estimate with its respective coverage interval and an improved neural network may be achieved by using the means of the joint PDF of the parameters

The proposed model can be used to reduce the frequency of jar tests by using the estimates of coagulant and alkalizer dosages. Both models can be useful to overcome the difficulty of determining chemical dosages in events such as heavy rain in order to be able to respond to significant changes in raw water quality to ensure the efficient operation of the water treatment plant.

ACKNOWLEDGEMENTS

Theauthorsthank FAPESB(processBOL1597/2008) for financial support, Clean Technology Network of Bahia (TECLIM) for general support and Mario Cezar Matosfor the fruitful discussions.

REFERENCES

Baughman, D. R., Neural Networks in Bioprocessing and Chemical Engineering, Academic Press(1995).

Baxter, S., Stanley, S.J., Zhang,Q., and Smith, D. W., Developing artificial neural network models of water treatment processes: a guide for utilities. Journal of Environmental Engineering and Science, 1(3), 201-211(2002).

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML (2008a). Evaluation of measurement data — Guide to the expression of uncertainty in measurement 1sted. Joint Committee for Guides in Metrology - JCGM 100:2008.

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML (2008b). Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method 1st ed., Joint Committee for Guides in Metrology - JCGM 101:2008.

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML (2011). Evaluation of measurement data – Supplement 2 to the "Guide to the expression of uncertainty in measurement" – Models with any number of output quantities, Joint Committee for Guides in Metrology - JCGM 102:2011.

Chen, T., Martin, E., and Montague, G., Robust probabilistic PCA with missing data and contribution analysis for outlier detection.Computational Statistics and Data Analysis, 53(10), 3706-3716 (2009).

Daszykowski, M., Kaczmarek, K., Vander Heyden, Y. and Walczak, B. (2007).Robust statistic in data analysis – A review basic concepts.Chemometrics and Intelligent Laboratory Systems, 85(2), 203-219 (2007).

Di Bernardo, L., Sabogal Paz, L.P. (2008). Selection of Water Treatment Technologies 1.LDIBE LTDA. São Carlos, Brazil.(in Portuguese).

Dobias, B., and Stechemesser, H., Coagulation and flocculation: theory and applications, 2nd edition. Taylor & Francis Group. Boca Raton (2005).

Dorea, C. C., Coagulant-based emergency water treatment. Desalination, 248(1-3), 83-90 (2009).

Feng, C-X. J., Yu, Z-G.S., Kingi, U., and Baig, M.P. (2005).Threefold vs. fivefold cross validation in one-hidden-layer and two-hidden-layer predictive neural network modeling of machining surface roughness data. Journal of Manufacturing Systems, 24(2), 93-107 (2005).

Flaten, T. P., Aluminium as a risk factor in Alzheimer's disease, with emphasis on drinking water. Brain Research Bulletin, 55(2), 187–196 (2001).

Hao, Y., and Wilamowski, B. M. (2011). Levenberg-Marquardt Training. Industrial Electronics Handbook, 2nd Edition, 5, 12-1 to 12-15 (2011).

Harmel, R.D., and Smith, P.K. (2007). Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. Journal of Hydrology, 337(3-4), 326-336 (2007).

Heddam, S., Abdelmalek, B., and Dechemi, N., Applications of radial-basis function and generalized regression neural networks for modeling of coagulant dosage in drinking water-treatment plant: Comparative study. Journal of Environmental Engineering, 137(12), 1209-1214 (2011).

Heddam, S., Bernard, A., and Dechemi, N., ANFIS-based modelling for coagulant dosage in drinking water treatment plant: a case study. Environmental Monitoring and Assessment, 184(4), 1953-1971 (2012).

Heddam, S., and Dechem, N., A new approach based on the dynamic evolving neural-fuzzy inference system (DENFIS) for modeling coagulat dosage (Dos): Case study of water treatment plant of Algeria. Desalination and Water Treatment, 53(4), 1045-1053 (2015).

Hernandez, H., and Le Lann, M.-V., Development of a neural sensor for on-line prediction of coagulant dosage in a potable water treatment plant in the way of its diagnosis. J. S. Sichman et al (Eds.) IBERAMIA-SBIA, LNAI 4140, 249-257. Springer-Verlag Berlin Keideberg(2006).

Hogg, R.V., and Craig, A.T., Introduction to mathematical statistics. Macmillan Publishing Co., Inc, New York (1998).

Johnson, R.A., Wichern, D.W., Applied Multivariate Statistical Analysis, 6th edition. Pearson Prentice Hall. USA (2007).

Joo, D.-S., Choi, D.-J., and Park, H., The effects of data preprocessing in the determination of coagulant dosing rate. Water Research 34(13), 3295-3302 (2000).

Lamrini, B., Benhammou, A., Le Lann, M.-V. and Karama, A., A neural software sensor for online prediction of coagulant dosage in a drinking water treatment plant. Transactions of the Institute of Measurement and Control, 27(3), 195-213 (2005).

Lingireddy, S.; and Brion, G. M. (Editors), Artificial Neural Networks in Water Supply Engineering. ASCE, American Society of Civil Engineers Press. Virgínia, USA (2005).

Lira, I., Monte Carlo evaluation of the uncertainty associated with the construction and use of a fitted curve. Measurement, 44(10), 2156–2164 (2011)..

Maier, H.R., and Dandy, G.C., 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 (2000).

Maier, H.R., Morgan, N., and Chow, C.W., Use of neural networks for predicting optimal alum doses and treated water quality parameters. Environmental Modelling & Software, 19(5), 485-494 (2004).

Montgomery, D.C., and Runger, G.C., Applied statistics and probability to engineers. John Wiley and Sons, New York (2015).

Oliveira-Esquerre, K.P., Kiperstok, A., Kalid, R., Sales, E., Teixeira, L., and Pires, V.M., Water and wastewater management in a petrochemical raw material industry. Computer Aided Chemical Engineering, 27, 1047-1052 (2009).

Oliveira-Esquerre, K.P., Seborg, D.E., Mori, M., and Bruns, R.E., Application of stead-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper Mill Part II. Nonlinear approaches. Chemical Engineering Journal, 105(1-2), 61-69 (2004).

Ouyang, Y., Nkedi-Kizza, P., Wu, Q.T., Shinde, D., and Huang, C.H., Assessment of seasonal variations in surface water quality. Water Research, 40(20), 3800-3810 (2006).

Robenson, A., Shukor, S.A., and Aziz, N., Development of process inverse neural network model to determine the required alum dosage at Segama water treatment plant Sabah, Malaysia. Computer Aided Chemical Engineering, 27, 525-530 (2009).

Rousseau, P. J., Debruyne, M., Engelen, S., and Hubert, M., Robustness and Outlier Detection in Chemometrics. Critical Reviews In Analytical Chemistry, 36(3-4), 221-242 (2006).

Schwaab, M. and Pinto, J. C., Análise de Dados Experimentais I. Fundamentos de Estatística e Estimação de Parâmetros. Rio de Janeiro: E-papers (2007).

Shrestha, D.L., Kayastha, N., and Solomatine, D.P., A novel approach to parameter uncertainty analysis of hydrological models using neural networks. Hydrology and Earth System Sciences Discussions, 13(7), 1235-1248 (2009).

Singh, I.B., Geological Evolution of Ganga Plain—An Overview. Journal of the Palaeontological Society of India, 41, 99-137 (1996).

Srivastav, R.K., Sudheer, K.P., and Chaubey, I., A simplified approach to quantifying predictive and parametric uncertainty in artificial neural network hydrologic models. Water Resources Research, 43(10), 1-12 (2007)

Willmott, C.J., and Matsuura, K., Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), 79-82 (2005).

Wu, G.-D., and Lo, S.-L., 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 (2008).

Wu, G-D., and Lo, S-L., Effects of data normalization and inherent-factor on decision of optimal coagulant dosage in water treatment by artificial network. Expert Systems with Applications, 37(7), 4974-4983 (2010).

Yan, M., Wang, D., Yu, J., Ni, J., Edwards, M., and Qu, J., Enhanced coagulation with polialuminum chlorides: Role o pH/Alkalinity and speciation. Chemosphere, 71(9), 1665-1673 (2008).

Yu, R.-F., Kang, S.-F., Liaw, S.-L., and Chen, M.-C., Application of artificial neural network to control the coagulant dosing in water treatment plant. Water Science and Technology, 42(3-4), 403-408 (2000).

Zhang, K., Achari, G., Li, H., Zargar, A., and Sadiq, R., Machine learning approaches to predict coagulant dosage in water treatment plants. Int. J. Assur. Eng. Manag., 4(2), 205-214 (2013).

Zhang, Q., and Stanley, S.J., Forecasting raw-water quality parameters for the North Saskatchewan River by neural network modeling. Water Res., 31(9), 2340-2350 (1997).