Prediction of organic matter removal from pulp and paper mill wastewater using artificial neural network

Jácina T. G. Morais[1], Karla P. Oliveira-Esquerre[2], Asher Kiperstok[3], Luciano M. Queiroz[3]

[1]Department of Chemical Engineering, Federal University of the Semi-Arid Region, Caraúbas, Rio Grande do Norte, Postcode - 59780000, Brazil

[2]Department of Chemical Engineering, Federal University of Bahia, Salvador, Bahia, Postcode - 40210630, Brazil

[3]Department of Environmental Engineering, Federal University of Bahia, Salvador, Bahia, Postcode - 40210630, Brazil

Corresponding author: Luciano M. Queiroz
e-mail: lmqueiroz@ufba.br
Phone: (+55) 71 32839796
Fax: (+55) 71 32839454

Prediction of organic matter removal from pulp and paper mill wastewater using artificial neural network

Abstract: The main purpose was to evaluate the application of Principal Component Analysis (PCA) as a preprocessing technique of the input data of a Multilayer Perceptron Model (MLP). The objective of the model was the prediction of organic matter removal from pulp and paper mill wastewater. Methods: The original data base covered a period of 1 427 consecutive days and contains the most frequently measured parameters. Two models (M1 and M2) were constructed without the application of PCA technique, and three (M3, M4 and M5) applying PCA to select principal components, discard original variables and exclude possible outliers. The data from each set were randomized and divided into three sets (training, validation and testing) each one containing 70%, 20% and 10%, respectively. The training algorithm was the Levenberg-Marquardt which is an adaptation of the back-propagation algorithm. The learning rate was 0.05 and the evaluation criteria used were the mean square error (RMSE) and the linear correlation coefficient (R^2). Results: PCA allowed discarding original variable and improving neural network performance without any loss of information. It was observed a marked difference in the predictive performance when the organic matter load was used as input (kg.day-¹). The model M4, which was built discarding two variables (pH and EC), proved to be the most suitable and the simplest model obtained. Conclusions: the choice of the best neural network model should not be done indiscriminately and carelessly. It is necessary to use various statistical parameters and perform comparison of models with different sizes and structures.
Keywords: artificial neural network, organic matter, principal component analysis, pulp and paper mill wastewater.

1. INTRODUCTION

Safer operation and control of industrial processes can be achieved by developing a modeling tool for predicting the plant performance, based on past observations of certain key product quality parameters. Performance assessment and monitoring of biological wastewater treatment process are usually made by collecting samples, and conducting physical and chemical analysis with daily attendance, what leads to an increase on the overall costs of the process. Besides, the numerical modeling to quantify the efficiency of contaminants removal, especially organic matter, is based on models whose kinetic constants are obtained often by studies with isolated cultures of microorganisms fed with specific substrate under laboratory scale. Microbial diversity and variability of organic substrate supplied to the microorganisms, associated with the variation of operating conditions in industrial process may limit the use of specific kinetic models for predicting performance of wastewater treatment systems. Therefore, since it is provided with a series of monitoring data, the application of predictive statistical tools is an attractive alternative that can provide information and correlations between industrial processes, wastewater characteristics and efficiencies of the wastewater treatment processes.

Nevertheless, some processes, such as industrial wastewater treatment, exhibit non-linear behaviors which are difficult to describe by linear mathematical models. However, the use of predictive models based on Artificial Neural Networks (ANN) to improve the operational control of wastewater treatment plants have been suggested in literature, and Multilayer Perceptron (MLP) has been successfully used [1-5]. Grieu *et al*. [6] presented a prediction procedure based on a MLP network to obtain influent and effluent organic matter concentrations. These authors indicated that neural modeling can be a useful tool to minimize operation costs and provide stability to the treatment process.

ANN normally relies on representative historical data of the process. Therefore, data preparation is an essential step for enhanced performance of predictive models. This task requires a careful analysis of the data, in order to define which variables best represent the system. Frequently, researchers face themselves with a large set of independent variables for possible inclusion in a multivariate analysis. In most cases, the inclusion of all variables in modeling is unnecessary and would be a serious obstacle to the correct interpretation of the data.

PCA is a multivariate statistical technique that reduces a complex system of correlations to a smaller number of dimensions. The main purpose is to reduce the dimensionality of a data set, consisting of a large number of interrelated variables, while retaining as much as possible of the variation, present in the data set [7-8]. Consider a data matrix **X** with **n** rows (observations) and **p** columns (variables). PCs are obtained by the diagonalization of the covariance matrix $\mathbf{X^TX}$, where $\mathbf{X^T}$ is a transposed matrix of **X**. The elements of the eigenvectors, called *loadings* (weights) in PCA terminology, represent the cosine directors, in other words, they express a contribution of each original axis in the new axis, the so called principal components (PCs). The eigenvalues represent the amount of variance described by the original eigenvectors.

There are several methods for selecting variables by PCA [9-12]. Jolliffe [13] proposed a method, designated B4, to discard original variables based on the loading vectors of the first PCs. In fact, the author tested five methods: a multiple correlation method, two principal component methods and two clustering methods. The methods were compared and all showed to be satisfactory for real, as well as artificial data, although none is shown to be significantly superior to the others. The principal component methods were more successful at producing best subsets. Thus, as PCA is useful for reduction of the variables and also to exclude outliers, this was the method chosen to improve network performance. Other research that has successfully applied PCA together with ANN is described by several studies [14-17].
The purpose of this work was to evaluate the application of PCA, as a preprocessing technique of the input data to select variables and PCs, and also to identify outliers, in order to obtain a prediction of organic matter removal from pulp and paper mill wastewater by a MLP model.

2. METHODS

2.1 Process description

The wastewater treatment system consists of two parallel tanks with mixing, and flocculation chambers to enhance particle flocculation, followed by an aerated lagoon system.

2.2 Data collection for prediction model

The original data base covered a period of 1 427 consecutive days, about 4-year daily record. In order to minimize loss of information due to exclusion of samples that contained high incidence of missing values (> 50%), the data sets contained only the most frequently measured variables: flow rate (Q), influent organic matter (COD_{in}), pH value, color, temperature, electrical conductivity (EC), wastewater flow coming from the pulp production (Q_{pulp}) and wastewater flow coming from the paper production (Q_{paper}). Biochemical oxygen demand (BOD) was not chosen as input variable because of the significant time of measurement, about five days, which made it impractical to build the model. Due to exclusion of sample that contains missing data, the exclusion of BOD and probable errors of measurement, the data set was reduced to 786 samples. Table 1 shows the basic statistical properties for the selected variables.

Table 1: Basic statistical properties of the selected variables.

Parameters	Mean	Standard deviation	Minimum	Maximum	Missing data (%)
Q ($m^3.day^{-1}$)	67 363.8	11 588.5	4 474	97 850	0
COD_{in}(mg $O_2.L^{-1}$)	561.5	104.2	136	925	6.2
pH	7.5	1.2	1.0	12.5	3.7
Color (unitsPt-Co)	464.4	123.6	41	1 317	3.6
Temperature (°C)	45.5	3.1	28	50.5	32.6
EC ($\mu S.cm^{-1}$)	1 530.9	378.1	379	5 810	3.9
Q_{pulp}(ton.day^{-1})	886.1	155.2	0	1 112.1	7.9
Q_{paper}(ton. day^{-1})	1 042.7	94.2	382.4	1 304.8	6.5
COD_{out}(mg $O_2.L^{-1}$)	315.5	2.0	105	865	5.8

Five models were constructed to predict the content of organic matter in the effluent of the aerated lagoon (COD_{out}). Model 1 (M1) was constructed quantifying the organic matter present in the wastewater as concentration of COD (mg $O_2.L^{-1}$) while the organic load (COD_{load}), calculated by the multiplication of the COD concentration and flow rate, was used in Models 2 to 5. PCA was applied to reduce the dimensionality of data set, in order to select PCs original variables and exclude possible outliers in Models 3 to 5.

2.3 Artificial neural network structure

The B4 method was used to discard original variables, based on the weight vectors of the first principal component. The procedure began by finding the original variable that had the highest absolute weight on the first PC. Then, this variable was placed in the selected set. The method continued by inspecting the weights of the original variables on the second PC. Once more, the variable with the highest absolute weight was selected and placed in the set (unless it was already selected; in which case, the variable with the next highest absolute weight must be selected). This procedure was repeated until the most-important PCs were checked (eingevalue > 0.7).
Multilayer perceptron (MLP) was the artificial neural network used for the prediction of the amount of organic matter effluent of the aerated lagoon (COD_{out}). The training algorithm was the Levenberg-Marquardt, which is an adaptation of the backpropagation algorithm. This algorithm is normally used for

ANN, with a little or moderate training set (up to several hundred weights) since it requires a large storage memory for execution. It has been proven to be fast, convergent and robust [1].
The neural network parameters can be changed to reach the suitable network architecture, aiming to find a model with a more satisfactory result. The network parameters that changed on the length of the training were: learning rate, number of hidden layers and number of neurons per each hidden layer. The data set were randomized and divided into three sets: training, validation and test. The transfer functions were log-sigmoid and linear for the intermediate and output layer, respectively.
The linear activation function for the output neuron was appropriate for continuous-variable targets. Sigmoidal activation functions for the input and hidden neurons were needed to introduce nonlinearity into the network. Without nonlinearity, hidden layers would not make the nets any more powerful than plain perceptrons (which do not have any hidden units, just input and output units). Sigmoidal activation functions are usually preferable to threshold activation functions [18].

2.4 Evaluation of the ANN model performance

The performance of each network model was evaluated by computing the mean square error (MSE), the linear correlation index (R^2) and adjusted linear correlation index (adjusted R^2). In contrast to R^2, the adjusted R^2 only increases if the additional model parameters significantly improve the regression results, to compensate the increase in regression degrees of freedom. Therefore, there is no similar statistical parameter to perform reliable comparative analyses of the predictive performances of ANN models than the adjusted R^2. The Minitab® and Matlab® were used to statistical analysis, PCA and ANN modeling, respectively.

3. RESULTS AND DISCUSSION

Table 2 shows the results of the variance and weights of the principal components. According to the criteria of the B4 method, the first five components should be preserved to build the model, since these PCs express 89.8% of the total preserved variance of the system. The most important variables were: flow rate (Q), COD_{load}, Q_{paper}, color, Q_{pulp}, and temperature, respectively. Since $COD_{load} = f(COD,Q)$ and Q present the same loading value, only COD_{load} was maintained as an input variable for the modeling. We highlight that the select variables, by B4 method, were the same when using training, validation or test data sets.

Table 2: Variance and weights of the principal components.

Principal Components	Variance	Explained variance (%)	Accumulated variance (%)
PC_1	2.7	33.8	33.8
PC_2	1.7	20.6	54.4
PC_3	1.3	15.7	70.1
PC_4	0.8	10.6	80.7
PC_5	0.7	9.0	89.7
PC_6	0.4	5.5	95.2
PC_7	0.4	4.8	100
PC_8	0	0	100

	Weights							
Principal Components	Q	COD_{load}	pH	Color	T	EC	Q_{pulp}	Q_{paper}
PC_1	*-0.49*	*-0.49*	0.27	-0.17	-0.28	0.38	-0.37	-0.24
PC_2	-0.42	-0.42	-0.35	-0.05	0.37	0.07	0.39	*0.48*
PC_3	0.06	0.06	0.53	*0.58*	0.32	0.49	0.16	0.10
PC_4	-0.11	-0.11	-0.26	0.70	-0.35	-0.22	*-0.39*	0.31
PC_5	-0.15	-0.15	-0.18	0.27	*0.58*	-0.29	-0.17	-0.64

The best results were obtained when the ANNs were composed by only one hidden layer, the learning rate was equal to 0.05 and the division of the data to perform training sets, validation and testing were equal to 70%, 20% and 10%, respectively.
Table 3 shows the results of the evaluation of the performance of each model. Considering the values of R^2 and $R^2_{adjusted}$, it can be noted that only M1 presented poor performance. A better performance was found when regarding models M2 to M5, once they were built considering the amount of influent organic matter expressed in load terms ($kg.day^{-1}$) instead of concentration terms. The Model M1 was built considering the concentration of organic matter as COD and the flow rate. Therefore, the fluctuation of values from these parameters may have been the cause of M1's poor performance. Considering the models M2 to M5, no significant differences were identified in the R^2 values. The number of iterations

varied from model to model; nevertheless this was not significant in this case, as it does not take more than 10 seconds to run each model.

Table 3: Evaluation of the models to predict the COD_{out} of the aerated lagoon.

Comparative parameters	Models M1 [1]	M2 [2]	M3 [2]	M4 [2]	M5 [2]
Inputs	8 original variables	8 original variables	5 PCs	5 original variable	8 original variables
Training data	706	706	706	706	706
Test data	80	80	80	80	80
Number of hidden neurons	1	1	1	1	1
Number of parameters	9	9	6	6	9
Transfer function	Sigmoid	Sigmoid	Sigmoid	Sigmoid	Sigmoid
Processing method	Back-propagation	Back-propagation	Back-propagation	Back-propagation	Back-propagation
Number of iterations	11	103	18	93	103
MSE test	2.3E-03	4.59E-08	1.9E-05	2.9E-08	4.0E-05
R^2 test	0.4508	0.9999	0.9953	0.9999	0.9999
Adjusted R^2	0.3753	0.9999	0.9796	0.9999	0.9999

[1] COD (mg $O_2.L^{-1}$) and [2] COD_{load} (kg.day^{-1})

The adjusted R^2 values calculated with the data from the models M1 to M5 also showed no significant difference, which means that the network performance was unaffected by the reduction in adjustable parameters. Therefore, the complexity of the models can be reduced by appropriate data preparation. In other words, the M3 presented a faster learning (18 interactions) when comparing the structure of the variability of the ANN models, using the original variables (M2) and the corresponding PCs (M3).This was expected since the M3 was built with a smaller number of input data, but the results also showed no significant loss of information, which can be considered an advantage of using PCs. Furthermore, this performance was repeated when we selected variables by B4 method (M4). Therefore, we conclude that the signal-to-noise ratio does not affect the PCA application in this case.

The error rate is generally more significant because it is a supervised neural network type. It should be noted that the model M3 was built considering only five PCs as predictor variables in ANN, but required information from eight original variables. The results obtained using PCA; excluding possible outliers (M5) were similar to M3, which means that the exclusion of outliers is unnecessary in this case. However, this result cannot be generalized. In fact, the presence of outliers can provide incorrect or misleading results, mainly during the construction of empirical models. The model M4 was built discarding pH and EC variables so, it was the most synthetic and simplest model obtained; therefore it was possible to conclude that the two variables discarded do not add information, nor influence the performance of the prediction model.

Table 4 shows the synaptic weights related to each input variable of the models. The results indicate that the synaptic weights of the variables influent flow rate and COD were significant to build the models M2, M4 and M5. On the other hand, the models M1 and M3 presented the weights of all variables in the same order of magnitude. This results indicate that some information, such as the COD concentration and the use of PCs as input variables of the MLP, were satisfactory for predicting the organic matter effluent of the aerated lagoon. However, for a better understanding of synaptic weights, it is necessary to perform a sensitivity analysis, which was not performed in this work.

The equation (1) was obtained from the model M4 and provides the prediction of removal of the organic matter from pulp and paper mill wastewater.

$$COD_{out} = \frac{8.67}{1 + e^{-0.46\,COD_{in} - 0.003\,Color - 0.0002\,Q_{paper} - 0.001\,Q_{pulp} + 0.0017}} - 4.3 \qquad \text{eq. (1)}$$

Table 4: Results of the synaptic weights of the models.

		Inputs variable of the MLP							
Models	COD	Q	COD	pH	Color	T	EC	Q_{Pulp}	Q_{Paper}
M1	$mg.L^{-1}$	-0.1514	-1.0645	0.1975	-0.0483	-0.0391	-0.3931	-0.0707	-0.0647
M2	$kg.day^{-1}$	-0.1243	-0.3201	0.0001	0.0003	0.0001	-0.0003	0.0002	0.0001
M4	$kg.day^{-1}$	-	-0.46	-	-0.003	0	-	-0.0002	-0.0001
M5	$kg.day^{-1}$	-0.1243	-0.3201	0.0001	0.0003	0.0001	-0.0003	0.0002	0.0001
		PC_1	PC_2	PC_3	PC_4	PC_5			
M3	$kg.day^{-1}$	-1.2451	-0.6821	0.1027	-0.1672	-0.1703	-	-	-

Figure 1 shows the comparison between predicted and measured values of the M4 test set. It is possible to observe that this model perfectly reproduces the overall variation observed in the biological treatment. Figure 2 shows the time series plots of the standardized residuals of the M4 model. It is noted that 95% of residuals fall in the range (-2, +2), which ensures the normality of residuals and the adequacy of the model. Standardization was carried out with the standard deviation of the validation data set.

Fig. 1 Time series plot of measured and predicted COD_{out} of the M4 model

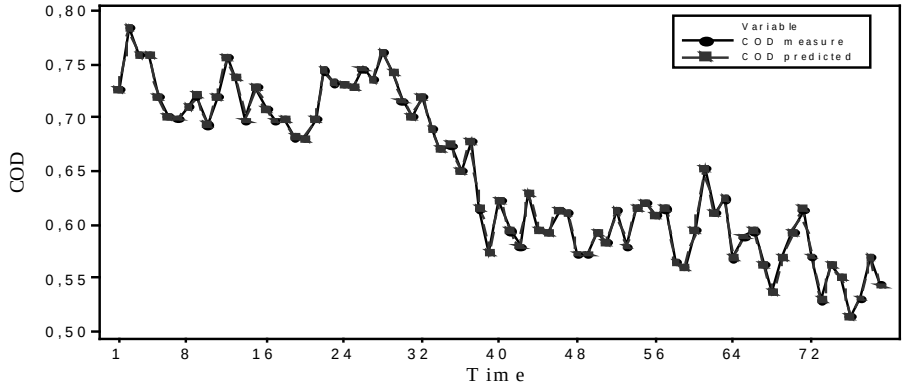
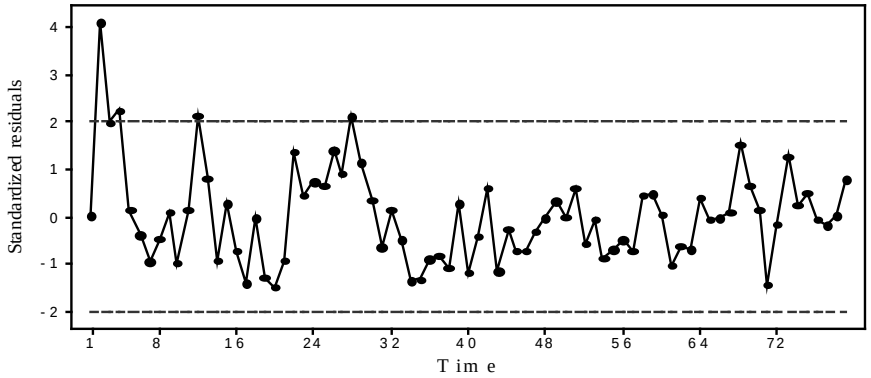


Fig. 2 Time series plot of the standardized residuals of the M4 model



4. CONCLUSIONS

This research investigated the use of Principal Component Analysis (PCA) as data-preprocessing techniques to build an artificial neural network model that allow the prediction of organic matter removal from a pulp and paper mill wastewater. It was concluded that the principal component analysis (PCA), applied to select input variables can be useful in neural network learning processes. The use of this technique allowed to reduce the number of parameters to be adjusted, without changing the performance of the model. The application of the PCA to discard original variable made it possible to improve neural network performance without any loss of information. The use of an ANN model may reduce costs by discarding unnecessary measurements. However, in this particular case, the PCA technique was unnecessary for outlier exclusion.

It is important to highlight that the choice of the best ANN model should not be done indiscriminately and carelessly. It is necessary to use various statistical parameters to assist the choice and the comparison of models of different sizes and structures. It is strongly recommended that the preprocessing data, in order to be meaningful, must be accompanied by a professional, who has expertise in the process.

ACKNOWLEDGMENTS
The authors would like to thank the Foundation for Research Support of the Bahia State for financial and material support.

REFERENCES
1. Hamed, M.M., Khalafallah, M.G., Hassanien, E.A.: Prediction of wastewater treatment plant performance using artificial neural networks. Environ. Modell. Softwe. 19, 919 – 928 (2004).
2. Mjalli, F.S., Al-Asheh, S., Alfadala, H.E.: Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. J. Environ. Manage. 83, 329 – 338 (2007).
3. Chen, H., Ning, S., Yu, R., Hung, M.: Optimizing the Monitoring Strategy of Wastewater Treatment Plants by Multi objective Neural Networks Approach. Environment Assess. 125, 325 – 332 (2007).
4. Akratos, C.S., Papaspyros, J.N.E., Tsihrintzis, V.A.: An artificial neural network model and design equations for BOD and COD removal prediction in horizontal subsurface flow constructed wetlands. Chem. Eng. J. 143, 96 – 110 (2008).
5. May, D.B., Sivakumar, M.: Prediction of urban stormwater quality using artificial neural networks. Environ. Modell. Softwe. 24, 296 – 302 (2009).
6. Grieu, S., Traore, A., Polit, M., Colprim, J.: Prediction of parameters characterizing the state of a pollution removal biologic process. Eng. Appl. Artif. Intele. 18, 559 – 573 (2005).
7. Kourti, T., Nomikos, P., MacGregor, J.F.: Process analysis, monitoring and diagnosis using multivariate projection methods. Chemo. and Intell. Lab. Sys. 28, 3 - 21 (1995).
8. Silva, A.P.D.: Efficient Variable Screening for Multivariate Analysis. J. Multivariate Anal. 76, 35 - 62 (2001).
9. Cumming, J.A., Wooff, D.A.: Dimension reduction via principal variables. Comput. Stat. Data An. 52, 550 – 565 (2007).
10. Cadima, J., Cerdeira, J.O., Minhoto, M.: Computational aspects of algorithms for variable selection in the context of principal components. Comput. Stat. Data An. 47, 225 – 236 (2008).
11. Fueda, K., Iizuka, M., Mori, Y.: Variable selection in multivariate methods using global score estimation. Comput. Stat. 24, 127 – 144 (2009).
12. Matteau, M., Assani, A. A., Mesfioui, M.: Application of multivariate statistical analysis methods to the dam hydrological impact studies. J. Hydrol. 371, 120 – 128 (2009).
13. Jolliffe, I.T.: Discarding Variables in a Principal Component Analysis. I: Artificial Data. Appl. Stat-J. Roy. St. C. 2, 160 - 173(1972).
14. Abbaspour, A., Baramakeh, L.: Application of principle component analysis artificial neural network for simultaneous determination of zirconium and hafnium in real samples. Spectrochim. Acta. 64, 477 – 482 (2006).
15. Bellamine, F.H., Elkamel, A.: Model order reduction using neural network principal component analysis and generalized dimensional analysis. Eng. Comput. 25, 443 - 463(2008).
16. Wu, C.L., Chau, K.W., Fan, C.: Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. J. Hydrol. 389, 146 – 167 (2010).
17. Palit, M., Tudu, B., Bhattacharyya, N., Dutta, A., Dutta, P.K., Jana, A., Bandyopadhyay, R., Chatterjee, A.: Comparison of multivariate preprocessing techniques as applied to electronic tongue based pattern classification for black tea. Anal. Chim. Acta. 675, 8 – 15 (2010).
18. Saraswathi, R., Saseetharan, M. K., Suja, S.: ANN–based predictive model for performance evaluation of paper and pulp effluent treatment plant. Int. J. of Computer App. Tech. 45, 280 – 289 (2012).