FEDERAL UNIVERSITY OF BAHIA, BRAZIL
POLYTECHNIC SCHOOL
DEPARTMENT OF ENVIRONMENTAL ENGINEERING
RESEARCH GROUP ON SOLID WASTE AND WASTEWATER
RESEARCH NETWORK ON CLEAN TECHNOLOGIES
http://www.teclim.ufba.br

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

Authors:
Jácina T. G. Morais
Karla P. Oliveira-Esquerre
Asher Kiperstok
Luciano Matos Queiroz

Athens - Greece
May 2015

Introduction and Motivation

- Safer operation and control of processes can be achieved by developing a modeling based on past observations of certain key product quality parameters
- Microbial diversity and variability of organic substrate associated with the variation of operating conditions may limit the use of specific kinetic models for predicting performance of wastewater treatment systems
  - So, 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.

Introduction and Motivation

- Industrial wastewater treatment, exhibit non-linear behaviors which are difficult to describe by linear mathematical models
- There are several reports indicating the use of predictive models based on Artificial Neural Networks (ANN) could improve the operational control of wastewater treatment plants.
- Data preparation is an essential step for enhanced performance of predictive models. PCA is a multivariate statistical technique that reduces a complex system of correlations to a smaller number of dimensions.

Objective

- 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

Methods

- The original data base covered a period of 1,427 consecutive days, **about 4-year daily record**. 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**.

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 (µS.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

Methods

- Five models were constructed to predict the content of organic matter in the effluent of the aerated lagoon (COD_{out}):
- Model 1 (M1): organic matter was quantified as concentration of COD ($mg\ O_2.L^{-1}$)
- Models 2 to 5: organic matter was quantified as organic load (COD_{load}), calculated by the multiplication of the COD concentration and flow rate.
- 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.
- The B4 method was used to discard original variables, based on the weight vectors of the first principal component.

Methods

- 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.
- Sigmoidal activation functions for the input and hidden neurons were needed to introduce nonlinearity into the network.
- 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).
- The Minitab® and Matlab® were used to statistical analysis, PCA and ANN modeling, respectively.

Results

- 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.

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

Principal Components	Weights							
	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

Results

- 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.

	Models				
Comparative parameters	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

Results

- 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).
- 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.
- The model M4 was built discarding pH and EC variables so, it was the most synthetic and simplest model obtained

Results

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	-	-	-

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

Results

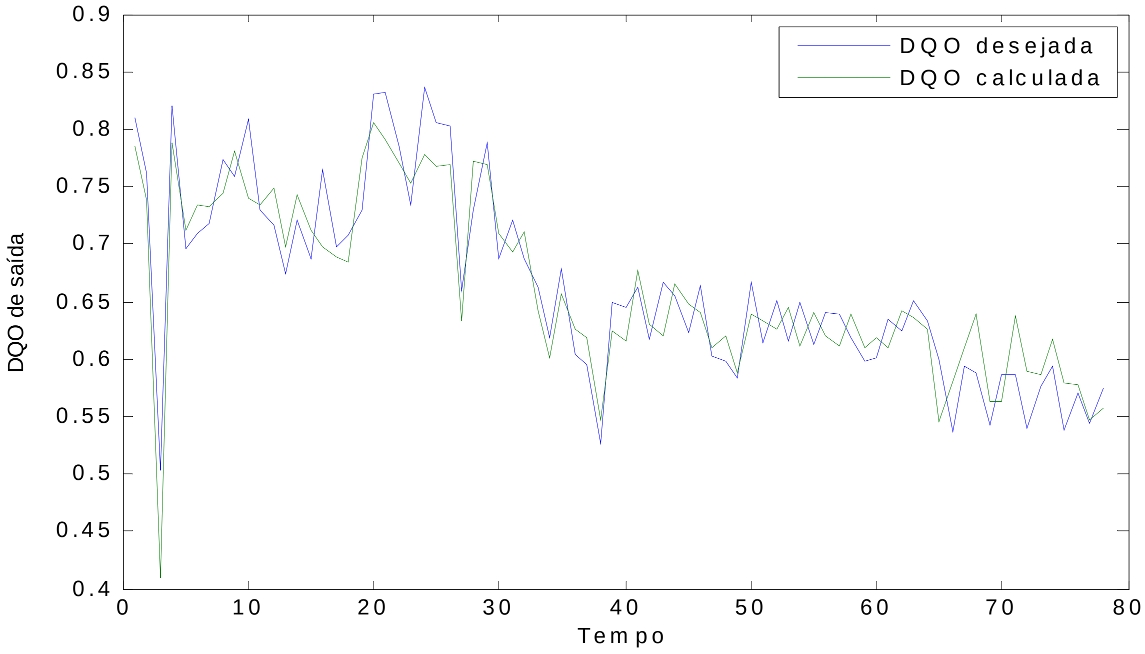

Figure 1 - Results of Model M4

IWWATV

Conclusions

- Principal component analysis (PCA), applied to select input variables can be useful in neural network learning processes.
- The application of the PCA to discard original variable made it possible to improve ANN performance without any loss of information. 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 strongly recommended that the preprocessing data, in order to be meaningful, must be accompanied by a professional, who has expertise in the process.

FEDERAL UNIVERSITY OF BAHIA, BRAZIL
POLYTECHNIC SCHOOL
DEPARTMENT OF ENVIRONMENTAL ENGINEERING
RESEARCH GROUP ON SOLID WASTE AND WASTEWATER
RESEARCH NETWORK ON CLEAN TECHNOLOGIES
http://www.teclim.ufba.br

THANK YOU FOR YOUR ATTENTION!!

Our contacts:

lmqueiroz@ufba.br
Phone: (+55) 71 32839796
http://www.teclim.ufba.br

IWWATV