JOURNAL OF CHEMOMETRICS
J. Chemometrics 2007; **21**: 156–164
Published online 18 July 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/cem.1054

Raw data pre-processing in the protozoa and metazoa identification by image analysis and multivariate statistical techniques

Y. P. Ginoris[1]*, A. L. Amaral[2,3], A. Nicolau[2], M. A. Z. Coelho[1] and E. C. Ferreira[2]

[1]Departamento de Engenharia Bioquímica, Escola de Química/UFRJ, Centro de Tecnologia, E-113, Cidade Universitária, Ilha do Fundão Rio de Janeiro, CEP 21949-900, Brazil
[2]Centro de Engenharia Biológica, Campus de Gualtar, Universidade do Minho, 4710-057 Braga, Portugal
[3]Departamento de Tecnologia Química e Biológica—ESTIG, Instituto Politécnico de Bragança, Apartado 1038, 5301-854 Bragança, Portugal

Received 13 October 2006; Revised 9 April 2007; Accepted 25 April 2007

Different protozoa and metazoa populations develop in the activated sludge wastewater treatment processes and are highly dependent on the operating conditions. In the current work the protozoa and metazoa groups and species most frequent in wastewater treatment plants were studied, mainly the flagellate, sarcodine, and ciliate protozoa as well as the rotifer, *gastrotrichia*, and *oligotrichia* metazoa. The work is centered on the survey of the wastewater treatment plant conditions by protozoa and metazoa population using image analysis, discriminant analysis (DA), and neural networks (NNs) techniques, and its main objective was set on the evaluation of the importance of raw data pre-processing techniques in the final results. The main pre-processing techniques herein studied were the raw parameters reduction set by a joint cross-correlation and decision trees (DTs) procedure and two data normalization techniques: logarithmic normalization and standard deviation normalization. Regarding the parameters reduction methodology, the use of a joint DTs and correlation analysis (CA) procedure resulted in 28 and 30% reductions in terms of the initial parameters set for the stalked and non-stalked microorganisms, respectively. Consequently, the use of the reduced parameters set has proven to be a suitable starting point for both the DA and NNs methodologies, although for the DA an initial logarithmic normalization step is advisable. For the NNs analysis a standard deviation normalization procedure could be considered for the non-stalked microorganisms regarding the operating parameters assessment. Copyright © 2007 John Wiley & Sons, Ltd.

KEYWORDS: protozoa; metazoa; image analysis; pattern recognition

1. INTRODUCTION

1.1. Protozoa and metazoa in activated sludge

The activated sludge wastewater treatment process relies on the activity of a bacterial aerobic culture suspended in the aerated tank fed with fresh effluent. The presence of such a bacterial culture allows also the development of a microfauna consisting mainly of predator organisms such as protozoa and metazoa. According to Madoni [1], and in order to take place an efficient treatment, there should be a high protozoa density ($>10^3$ per ml), dominant crawling and sessile forms and a well diversified community, with no overwhelming predominant species or group of species. When such is not the case, the dominant group or group's knowledge may give some clues for the wastewater treatment plant diagnosis, namely on the final effluent quality, aeration, sludge age, nitrification, and presence of toxic substances.

The exact number of protozoa species remains unknown, but over 50 000 species have been already identified so far [2,3]. In activated sludge, the three major protozoa groups are the flagellates, sarcodines, and ciliates [4], with a large predominance of ciliate species, that nourish mainly on bacteria, although some fed of other ciliates or flagellates (carnivorous). According to Madoni [1], bacterivore ciliates can be divided into three groups with respect to their feeding behavior: free swimming (moving freely in the effluent), crawling (grazing and living in the surface of the flocs), and sessile (attached to sludge flocs by a stalk structure). With respect to the metazoa the main groups present in a wastewater treatment plant are rotifers, nematodes, *gastrotrichia*, and *Oligotricihia* [4].

*Correspondence to: Y. P. Ginoris, Departamento de Engenharia Bioquímica, Escola de Química/UFRJ, Centro de Tecnologia, E-113, Cidade Universitária, Ilha do Fundão Rio de Janeiro, CEP 21949-900, Brazil.
E-mail: yovanka.perez@gmail.com

Different protozoa and metazoa populations develop in the activated sludge wastewater treatment processes and are highly dependent on the operating conditions. For instance, food availability will be decisive on the dominant group(s). Flagellates, sarcodines, and small free-swimming ciliates require a higher amount of bacteria due to their inefficient food capture ability. During the plant start-up, when there is a low hydraulic residence time (HRT) and a high food to microorganisms (F:M) ratio, these protozoa dominate. On the opposite, sessile ciliates and metazoa increase when there is a high HRT and a low F:M ratio due to their ability of floc adhesion or to their more efficient food capture mechanism [4]. Therefore, protozoa and metazoa populations are quite dependent on the sludge age which is in turn dependent on the plant organic load. Generally speaking, the colonization of a WWTP can be divided in three stages [1,5]: a first stage characterized by the presence of 'pioneer' species such as flagellates and free-swimming ciliates which are independent of the incoming raw effluent; a second stage of sludge formation when flagellates and free-swimming ciliates disappear progressively whereas sessile and crawling ciliates increase both in number as in species; a third stage where the population structure reflects the established conditions as a function of the balance between the organic load and the produced, removed, and recycled sludge.

In the current work the protozoa and metazoa groups and species most frequent in wastewater treatment plants were studied, mainly flagellate, sarcodine, and ciliate protozoa as well as rotifer, *gastrotrichia*, and *oligotrichia* metazoa, presented in Table I.

1.2. Image processing and multivariate statistical analysis

The major drawbacks on the use of protozoa and metazoa in WWTP diagnosis derive from the need of skilled workers specialized in zoology or protozoology, and that the identification task is time consuming. Image analysis emerges, then, as a potentially alternative tool to overcome such problems. However, up to date there have been few studies in this field such as the works of Amaral and collaborators [6], da Motta and collaborators [7], and Golz and Lange [8].

Table I. Protozoa and metazoa studied in this work

Protozoa	Flagellate		*Peranema* sp.
	Sarcodine		*Arcella* sp. *Euglypha* sp.
	Ciliate	Free swimming Carnivorous	*Trachelophyllum* spp. *Coleps* sp. *Litonotus* sp. Suctoria (sub-class)
		Crawling	*Aspidisca cicada* *Euplotes* sp. *Trithigmostoma* sp. *Trochilia* sp.
		Sessile	*Carchesium* sp. *Epistylis* spp. *Opercularia* sp. *V. convallaria* *V. aquadulcis* *V. microstoma* *Zoothamnium* sp.
Metazoa	Rotifer		Digononta (order) Monogononta (order)
	Gastrotrichia		Nematoda (sub-class)
	Oligotrichia		*Aelosoma* sp.

The objective of image processing and analysis methodology resides on obtaining a set of morphological descriptors representative of the protozoa and metazoa microorganisms. These descriptors may be subsequently studied and organized in a manner that allows the isolation and identification of each species, genus, order or sub-class by multivariate statistical techniques such as discriminant analysis (DA), neural networks (NNs), and decision trees (DTs).

1.2.1. *Discriminant analysis*

DA is a technique that determines new variables (discriminant functions) as linear combinations of the original descriptors, with the goal of increasing the inter-class variability and, thereby, obtains a better separation between the studied species and/or groups. Furthermore, in DA, the groups or classes of data are modeled with the aim of reclassifying the given object with a low error risk and of classifying new objects using the new discriminant functions [9]. The objects, coordinated in the new discriminant functions space, are obtained from the original descriptors.

1.2.2. *Neural networks*

An artificial NN is a biologically inspired computational model consisting on processing elements (neurons) operating in parallel and connections between them with associated coefficients (weights). Although a single neuron can perform certain simple information-processing functions, the power of NNs comes from connecting neurons in networks. This assembly, which is called the neuronal structure, is then trained with the help of recall algorithms. NNs can be adjusted, or trained, so that a particular input leads to a specific target output by the comparison of the network output and the target, until a match is obtained. There are three major learning paradigms, each corresponding to a particular abstract learning task: supervised learning (output values given), unsupervised learning (no output values given, usually used in statistical modeling, compression, filtering, blind source separation, and clustering), and reinforcement learning (control problems, games, and other sequential decision making tasks). An artificial NN can be defined by the following parameters [10]: type of neurons (nodes), connectionist architecture, training algorithm, and learning algorithm. The connectionist architecture is the organization of the connections between models and observes the NN number of layers and the nodes number in each layer.

1.2.3. *Decision trees*

A regression tree is a predictive model based on the ability to submit the input data matrix with a series of consecutive yes or no questions, and accurately predict a given response vector. Each question evaluates a given condition (either continuous or discrete) and, depending on the answer

DOI: 10.1002/cem

proceeds to a new question or arrives at the fitted response value.

However, one should be careful to avoid over fitting. In fact, a DT might be trained to fit so perfectly the data set that would not be appropriated for predicting new values. That is so when the tree has too much branches and the lower ones are strongly affected by outliers and other artefacts on the data set. One way to determine the best tree size is by cross-validation, which determines a resubstitution estimated by the error variance, leading to a series of pruned trees. Then the best tree is chosen as the tree presenting the residual variance that is no more than one standard error above the minimum value along the cross-validation line.

The present work is the follow-up of previous studies [6,11] on the survey of the wastewater treatment plant conditions by protozoa and metazoa population by image analysis, principal component analysis, DA, and NNs techniques, and its main objective was set on the evaluation of the importance of raw data pre-processing techniques in the final results. The main pre-processing techniques herein studied are the raw parameters reduction set by a joint cross-correlation and DTs procedure and two data normalization techniques: logarithmic normalization and standard deviation normalization.

2. MATERIALS AND METHODS

After the mixed liqueur collection, a drop of the sample was deposited carefully in a slide and covered with cover slip for visualization and image acquisition using a bright field microscope. A total of 22 different protozoa and metazoa was evaluated and the total magnification for acquiring each group was dependent on the microorganism size as follows: *Aelosoma* sp. (25 and 100 times); Nematoda (100 and 250 times); Digononta, Monogonta, *Arcella* sp., and *Euglypha* sp. (250 and 400 times); *Aspidisca cicada*, *Carchesium* sp., *Epistylis* spp., *Euplotes* sp. *Litonotus* sp., *Coleps* sp., *Opercularia* sp., *Peranema* sp., Suctoria, *Trachellophyllum* spp., *Trithigmostoma* sp., *Trochilia* sp., *V. aquadulcis*, *V. microstoma*, *Vorticella* sp., and *Zoothamnium* sp. (400 times). The dimensions of metric units (μm) were correlated with the corresponding pixel units using a micrometric slide.

Among the evaluated groups: two species of *Epistylis* and two species of *Trachellophylum* were additionally analyzed. Moreover, a group of microorganisms with similar morphological characteristics of *Epystilis* sp. and *Opercularia* sp. was included due to the fact that when these organisms occur with the closed buccal apparel it is quite difficult distinguish one group from the other. Finally, the frontal and lateral views of *Arcella* sp., *A. cicada*, and *Trithigmostoma* sp. were also analyzed, on cause of their axial lack of similitude.

Samples from two sites, Braga in Portugal and Nancy in France, were treated. The image acquisition system used in Nancy was composed by a *Leitz Dialux 20* optic microscope (Leitz, Wetzlar) coupled to a gray scale video camera *Hitachi CCTV HV-720E (F)* (Hitachi, Tokyo). The images were grabbed to the computer in 768×576 pixels and 8-bit format (256 gray levels) by a *Matrox Meteor* frame grabber (Matrox, Montreal) using the *Visilog 5* commercial software (Noesis, S.A., les Ulis). In Braga, the acquisition system was composed by an optic microscope *Zeiss Axioscop* (Zeiss, Oberkochen) coupled to a *Sony CCD ACV D5CE* gray scale video camera (Sony, Tokyo) and connected to a PC through the *Data Translation DT 3155* frame grabber (Data Translation, Marlboro), in order to convert the analogical voltage signal of the camera on an 8-bit digital 768×576 pixels matrix. This digital representation was then acquired, exhibited in the computer screen, and stored to the computer using the commercial software *Image-Pro® Plus* (Media Cybernetics, Silver Spring).

A smaller set of images was acquired during the present work using an acquisition system consisting of a *Leitz Laborlux S* optic microscope (Leitz, Wetzlar) coupled to a *Zeiss Axion Cam HR* video camera (Zeiss, Oberkochen). The images acquisition was performed in 1300×1030 pixels and 8-bit format through the commercial software *Axion Vision* 3.1 (Zeiss, Oberkochen).

2.1. Image analysis program

The procedure to process the acquired images and determine the morphological parameters, was adapted from the *ProtoRec* v.4 program previously developed by Amaral and collaborators [11] and converted to the *Matlab* 7.0 (The MathWorks, Inc., Natick) language.

The first step of the image analysis procedure consists on the gray-level images pre-treatment by applying a local histogram equalization to enhance the contrast of each region in the image, followed by the use of the median filter to perform a noise reduction and the *Bottom hat* filter to emphasize the organisms borders. The resulting images are then combined for a better differentiation between the organism's borders and the background. After the pre-treatment step, a polygonal region of interest (ROI) around the selected organism is defined by the user. Once defined the ROI, the image is segmented by thresholding the organism's borders, through a value defined either manually or automatically using Otsu's [12] or entropy methods [13].

In the subsequent stage debris material (small artefacts and other materials that may interfere with the analysis) is eliminated by a series of morphological operations applied to the binary images including morphological closing, filling, and opening operations. Figure 1 represents the main steps of the image analysis procedure and Figure 2 illustrates the schematic representation of ProtoRec program.

The determination of the protozoan and metazoan morphological parameters is performed in two stages. In the first stage, the parameters are computed to the whole organism's body including their external structures such as flagella, cilia, cirri, and stalk. In the second stage, the parameters are determined for the organism's body core, i. e., after the removal of all external structures. These descriptors were subsequently studied and organized in a manner that allowed the isolation and identification of each species, genus, order, or sub-class. Bearing this purpose in mind, the multivariate statistical techniques DA and NNs were performed using the *Matlab* 7.0 platform (The MathWorks, Inc., Natick).

Table II presents the morphological descriptors determined for both the whole organism's body and the organism's body without the external structures. Except

DOI: 10.1002/cem

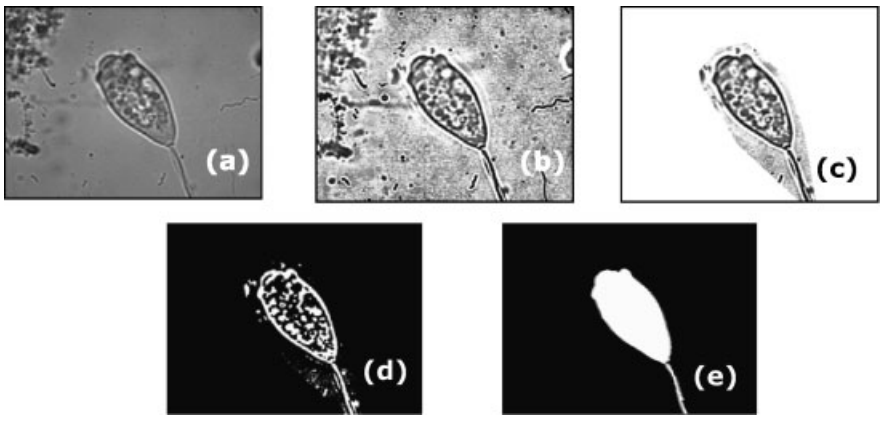

Figure 1. Main steps of the program: original image (a); pre-treated image (b); region of interest (c); binary image after segmentation (d); and final image (e).

when explicitly indicated, the morphological descriptors herein described were determined according to the *Matlab* built in functions.

Some descriptors were specifically designed for the protozoa and metazoa microorganisms, such as the mean body width versus body width ratio ($W_M W_B$), the mean stalk width versus mean body width ratio ($W_S W_{MB}$), and the mean stalk width (W_{Stk}). The stalk length S_{Stk} was determined by the following expression:

$$L_{Stk} = \frac{\frac{P_{Stk}}{2} + \sqrt{\left(\frac{P_{Stk}}{2}\right)^2 - 4A_{Stk}}}{2} \quad (1)$$

where P_{Stk} is the stalk perimeter.

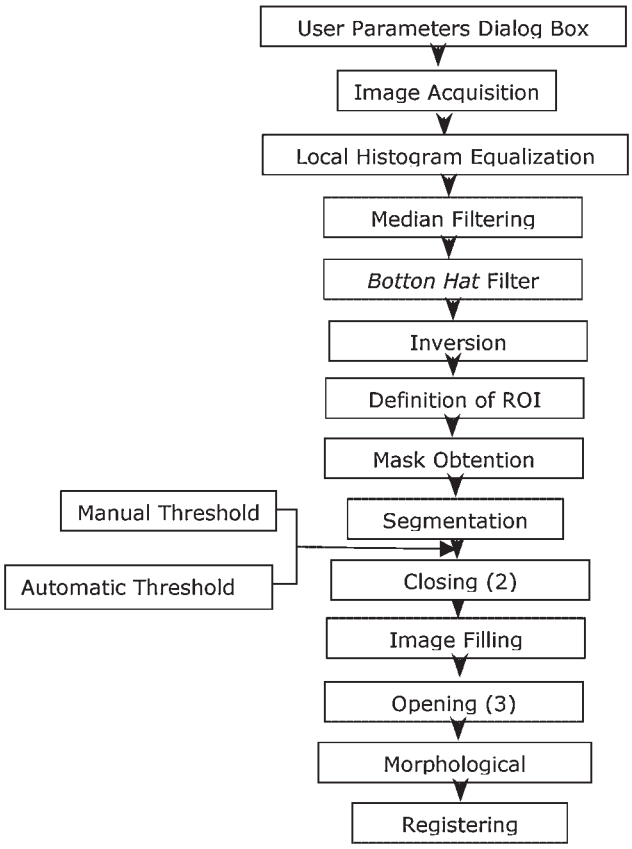

Figure 2. Schematic representation of ProtoRec program.

All of the descriptors were determined for the full protozoan and metazoan organism (including the external structures) as well as for the protozoan and metazoan body (without external structures), except for the mean stalk width and for $W_S W_{MB}$.

2.2. Multivariate statistical techniques

For the DA and NNs techniques the studied microorganisms were first separated into two easily recognizable classes: stalked and non-stalked microorganisms. This step is performed by the user to simplify and speed up the image analysis program since it represents a quite simple characteristic to

Table II. Morphological descriptors computed for protozoa and metazoa

Morphological descriptor	Mathematical expression
Surface (S)	*Matlab* built in
Equivalent diameter (D)	$\sqrt{(4A/\pi)}$
Perimeter (P)	*Matlab* built in
Length (L)	*Matlab* built in
Width (W)	*Matlab* built in
Mean width (W_M)	$W_M = S/L$
Feret factor (FrF)	$FrF = L/W$
Eccentricity (Ecc)	*Matlab* built in
Form factor (FF)	$FF = P^2/(4\pi S)$
Largest concavity index (LCI) [13]	—
Robustness (Rob) [13]	—
Concavity ratio (CR) [13]	—
Convexity (Conv)	$(Conv = P_{Conv}/P)^*$
Compactness (Comp)	$Comp = D/L$
Solidity (Sol)	$(Sol = S/S_{Conv})^{**}$
Euclidian distance map fractal dimension (D_{EDM}) [14]	—
Mass fractal dimension (D_{BM}) [15]	—
Surface fractal dimension (D_{BS}) [15]	—
Area vs. perimeter fractal dimension (D_{AvsP}) [16]	—
Mean body width vs. body width ratio ($W_M W_B$)	$W_M W_B = W_{MB}/W_B$
Mean stalk width vs. mean body width ratio ($W_S W_{MB}$)	$W_S W_{MB} = W_{Stk}/W_{MB}$
Mean stalk width (W_{Stk})	$(W_{Stk} = S_{Stk}/L_{Stk})^{***}$

* P_{Conv} is the convex envelope perimeter.
** S_{Conv} is the convex envelope surface.
*** S_{Stk} is the stalk surface and L_{Stk} the stalk length.

DOI: 10.1002/cem

Table III. Number of individual organisms present in the training set

acic	aelo	arce	carc	cole	digo	epis	ep/op	eugl	eupl	lito	mono
134	46	108	67	67	57	67/96	67	67	67	67	67

nema	oper	pera	suct	trac	trit	troc	vaqu	vcon	vmic	zoot
37	47	67	38	86	78	46	67	67	67	67

establish. Subsequently, DA and NNs were performed for the whole set of the microorganisms training set.

Initially, a training set of each 22 microorganisms was used for the determination of the discriminant functions and of the NN architecture. Regarding the stalked group two different analyses were performed: an analysis with the two *Epistylis* species as two different groups containing 10 groups and a second one with the two species represented in a single group in a total of 9 groups. For the non-stalked set a total of 18 groups were analyzed due to the fact that two different *Trachellophylum* species were studied and for *A. cicada*, *Arcella*, and *Trithigmostoma* species both front and side views (in separate groups) were treated. For validation purposes a different set of individuals (test set) of each 22 microorganisms was used with a third of individual organisms number of the training set and the same number of groups. The number of individual organisms used in each case is presented in Tables III and IV. The two values in the *Epistylis* column are reported to the cases where the two *Epistylis* species were analyzed as a single group or two different groups, respectively.

In this work the protozoa and metazoa are represented by: *A. cicada* (acic), *Aelosoma* sp. (aelo), *Arcella* sp. (arce), *Carchesium* sp. (carc), *Coleps* sp. (cole), Digononta order (digo), *Epistylis* spp. (epis), *Euglypha* sp. (eugl), *Euplotes* sp. (eupl), *Litonotus* sp. (lito), Monogononta order (mono), Nematoda sub-class (nema), *Opercularia* sp. (oper), *Peranema* sp. (pera), Suctoria sub-class (suct), *Trachelophyllum* spp. (trac), *Trithigmostoma* sp. (trit), *Trochilia* sp. (troc), *V. aquadulcis* (vaqu), *V. convallaria* (vcon), *V. microstoma* (vmic), and *Zoothamnium* sp. (zoot). When it was not possible to determine if a given organism was an *Epistylis* or an *Opercularia* (closed buccal apparel) the term ep/op was adopted.

2.2.1. *Discriminant analysis*

The performed DA was of a linear type, i.e., the multivariate normal (MVN) density function used was a relative log posterior density function (D) with a pooled estimate of variance. The value of the MVN density function was therefore determined for each individual organism regarding all the studied groups for both training and test sets.

Table IV. Number of individual organisms present in the test set

acic	aelo	arce	carc	cole	digo	epis	ep/op	eugl	eupl	lito	mono
66	23	54	33	33	29	33/47	33	33	33	33	33

nema	oper	pera	suct	trac	trit	troc	vaqu	vcon	vmic	zoot
20	23	33	18	43	39	22	33	33	33	33

In the validation process, and in order to determine each microorganism group, the MVN density function value was determined for all the individual organisms on the test set and for each group. Each organism was then assigned to the group where it presented the highest MVN density function value (D) provided that:

$$D < (\overline{D}_g - f\delta_g^D) \quad (2)$$

where $\overline{D}_g$ is the mean value of the MVN density function value for group g, δ_g^D is the standard deviation and f is a factor ranging from 0.25 to 5 in 0.25 step values. Microorganisms that do not comprise Equation (2) were classified as non-identified.

2.2.2. *Neural network*

The programmed NN was a two-layer (no hidden layers) feed forward NN with a back propagation algorithm and logistic sigmoidal activation functions. The Gradient Descent with momentum weight and bias learning function was the chosen back propagation learning function, whereas the mean squared error was used as the performance (error) function and its goal set to zero. For the stalked microorganisms two configurations (9/9 and 14/9) input/output nodes were tested when the two *Epistylis* species were analyzed as a single group and the 10/10 and 15/10 configurations for *Epistylis* as two different groups. Two back propagation training functions were used: Levenberg-Marquardt algorithm and the Resilient Backpropagation algorithm. Regarding the non-stalked microorganisms two other configurations (11/18 and 18/18) input/output nodes were used, and the back propagation training function was the Resilient Backpropagation algorithm. One hundred initial values for the NN architecture were tested for both the stalked and the non-stalked microorganisms and, for each a maximum of 500 epochs were computed.

In the validation process, the applied NNs aimed to obtain an output value of 1 for the microorganism correct group and 0 for all the other groups. Therefore, each microorganism was attributed to the group with a single higher output value larger than 0.01, and microorganisms with more than a single maximum group output were classified as non-identified.

2.3. Parameters reduction

The parameters reduction analysis was performed by a joint procedure of a DT to highlight the most important parameters and a correlation analysis (CA) to establish those parameters which presented less variability and therefore discard duplicate parameters. Both these techniques were carried out for the whole set of the parameters determined for the stalked (39 parameters) as well as the non-stalked (54 parameters) microorganisms, respectively. Therefore, the DT results allowed the selection of the most important parameters and the CA the exclusion of useless or duplicate parameters.

2.3.1. *Decision tree*

The chosen DT was a classification of text output data with Gini's Diversity Index as the split criterion and a minimum of

DOI: 10.1002/cem

10 observations in each impure node in order for that node to be split. Gini's Diversity Index (I_G) is based on the squared probabilities of membership of each target category in the node and reaches zero when all cases in the node are attributed to a single target category:

$$I_G(i) = 1 - \sum_{j=1}^{m} f(i,j)^2 \quad (3)$$

where $f(i,j)$ is the frequency of value j in node i.

In order to obtain the best pruned DT, the stop criterion was achieved by computing the best level using the test data as a test sample in which, applying the DT to this sample, a vector of cost values is returned. The best level chosen was the one that resulted in a lower value cost for the test data, when applied to the overall parameters and for the whole test data, for both the stalked and non-stalked analysis.

2.3.2. *Cross-correlation analysis*

A cross-correlation analysis was performed on the overall parameters for the whole training data and for both the stalked and non-stalked microorganisms. Therefore, the correlation among the 34 parameters for the stalked species training data and for the 54 parameters for the non-stalked training data was computed. Each pair of parameters, presenting a correlation higher than 0.9, was discarded regarding the choice of which parameter upon the relative importance to the data variability.

2.4. Normalization techniques

In order to normalize the results two different approaches were studied: logarithmic normalization and standard deviation normalization. Each procedure was applied to the stalked and non-stalked microorganisms training and test data, respectively. In the logarithmic normalization procedure, the natural logarithm was computed for each parameter, whereas for the standard deviation normalization the average and standard deviation values were calculated and the parameters values normalized according to:

$$X_{Norm} = \frac{(X - \overline{X})}{\delta_X} \quad (4)$$

where X_{Norm} is the normalized parameter value, $\overline{X}$ is the parameter average value for the stalked or non-stalked training set, and δ_X the parameter standard deviation value for the stalked or non-stalked training set.

3. RESULTS

The results obtained for the studied NN procedure allowed to determine small to negligible differences between the 18/18 and 11/18 non-stalked neural architectures although for the reduced and normalized results slight improvements in 18/18 architecture were observed. However, and given the higher computing speed in the 11/18 configuration it was considered that this architecture complies within these work objectives. With respect to the stalked microorganisms the configuration 15/10 led to better results and therefore, proved to have a real advantage over the 10/10 architecture.

Regarding the parameters reduction methodology, the use of joint DTs and CA procedure resulted in 28 and 30% reductions in terms of the initial parameters set for the stalked and non-stalked microorganisms, respectively. Therefore, 28 of the initial 39 parameters determined for the stalked identification and 38 of the initial 54 parameters determined for the non-stalked microorganisms identification were found to bear importance. With respect to the DTs technique it allowed to establish the importance of 16 parameters for the stalked species and 26 to the non-stalked ones. The performed CA allowed discarding 11 parameters for the stalked species and 16 for the non-stalked ones.

It was also studied in this work the analysis of parameters normalization in the final results. With that purpose two normalization techniques were studied: the logarithmic normalization and the standard deviation normalization. The comparison between the use of the complete parameters set and the reduced set as well as the two studied normalization techniques are presented in Tables V and VI.

Analyzing the results of the DA it can be found that for both the stalked and non-stalked microorganisms the parameters reduction did not present a significant effect, although the overall performance slightly decreased (less than 2.5% in both cases). With respect to the NNs and for both the stalked and non-stalked microorganisms the parameters reduction resulted in a slight increase (ranging from 1.8 to 3.2%) on the overall performance. It seems fair to withdraw that the reductions of 28% (stalked) and 30% (non-stalked) in the parameters set resulted in small to negligible effects in the results, and may, therefore, form the basis of future analyses.

From the analysis of the normalization results it seems clear that for the DA the logarithmic normalization presented

Table V. Recognition, misclassification, and overall DA and NN performance for the non-stalked microorganisms with the complete, reduced, and normalized set

		Rec. (%)	Misc. (%)	Overall (%)
DA	All	92.5	7.3	85.8
	Reduced	91.5	8.5	83.7
	Log normalized	92.5	7.5	85.6
	Standard normalized	91.7	8.1	84.3
NN	All	91.3	7.5	84.5
	Reduced	92.9	7.1	86.3
	Log normalized	93.3	6.5	87.3
	Standard normalized	93.5	5.5	88.4

Table VI. Recognition, misclassification, and overall DA and NN performance for the stalked microorganisms with the complete, reduced, and normalized set

		Rec. (%)	Misc. (%)	Overall (%)
DA	All	71.7	27.6	51.9
	Reduced	70.3	29.7	49.4
	Log normalized	71.7	28.3	51.4
	Standard normalized	69.6	30.4	48.4
NN	All	70.6	29.4	49.9
	Reduced	72.7	26.9	53.1
	Log normalized	70.6	29.4	49.9
	Standard normalized	69.6	30.4	48.4

DOI: 10.1002/cem

Table VII. Overall performance for the groups and ciliates identification, final effluent quality, aeration and nitrification assessment, and sludge age determination

		Groups	Ciliates	Effluent quality	Aeration	Nitrification	Age
DA	All	95.3	94.8	80.7	83.0	81.6	89.2
	Reduced	94.8	94.6	79.1	83.0	79.3	89.9
	Best normalization	95.4	94.7	81.2	85.4	81.2	91.7
NN	All	94.9	92.2	80.5	82.6	82.2	89.3
	Reduced	96.3	94.8	82.7	83.7	85.3	90.6
	Best normalization	95.9	95.4	82.9	83.8	83.8	91.0

slightly better results than the standard deviation normalization in both stalked and non-stalked microorganisms. However, it is also true that the differences can be considered small to marginal (below 3%) regarding each other results as well as with the non-normalized results above-mentioned for both cases. When compared to the reduced set results, improvements ranging from 1.9% (non-stalked) to 3% (stalked) are observed and, therefore, when working with reduced parameter sets the normalization procedure may be considered for the analysis.

For the NN approach, slightly better results are observed with the standard deviation than with the logarithmic normalization procedure for the non-stalked microorganisms and opposite results for the stalked microorganisms, but in all cases below 1.5%. Comparing these results with the ones obtained for the reduced non-normalized set, regarding the non-stalked microorganisms, almost similar results were obtained (up 2.1%) and slightly worst results (up to less 4.7%) for the stalked were obtained with the normalization procedure. Therefore, the normalization procedure for the NN results has given no advantages for the stalked microorganisms and is therefore considered unnecessary. However, for the non-stalked ones a small gain was obtained for the standard deviation normalization implying that this procedure can be considered in future works.

The overall performance for the DA and NNs regarding the groups and ciliates identification, final effluent quality, aeration and nitrification assessment, and sludge age determination are presented in Table VII for the entire data set, reduced data set, and best normalization. The best normalization procedures were: the logarithmic normalization in the DA technique for both stalked and non-stalked microorganisms, whilst for the NN the standard normalization for the non-stalked microorganisms and no normalization for the stalked ones were mostly appropriate.

With respect to the DA technique it could be found that the reduction on parameters set does not affect significantly the overall results for the operational parameters survey (below 2.3% differences), whereas the use of the most favorable normalization procedures led to a marginal gain in the results for operational parameters survey (up to 2.4% in aeration assessment) regarding the reduced set results. As far as the NNs technique is concerned, marginal gains were found for the reduction of the parameters set (up to 3.1%), and no noticeable gains were observed (from less 1.5% up to 0.6%) regarding the best normalization related to the reduced set results.

Table VIII. Overall performance for the critical conditions assessment

		Low effluent quality	Low aeration	Fresh sludge
DA	All	87.8	83.6	82.8
	Reduced	89.0	85.1	83.5
	Best normalization	92.7	89.6	88.3
NN	All	88.1	84.6	83.2
	Reduced	86.3	80.3	84.7
	Best normalization	89.4	85.4	85.2

As referred earlier, the 28% (stalked) and 30% (non-stalked) parameters reduction did not have a strong negative effect on the results as well as the logarithmic and standard deviation normalization procedures produced no major advantages, mainly in the NNs procedure. Therefore, for future works, it seems licit to infer that the use of the reduced parameters set raw data may be a suitable starting point for the DA and NN methodologies, although an initial normalization step could be considered, especially for the DA technique.

The overall performance of the DA and NNs for the critical conditions assessment (low effluent quality, low aeration, and fresh sludge) is presented in Table VIII.

From the analysis of Table VIII it could be withdrawn that the use of the reduced data set in DA allowed a minor improvement on the results (up to 1.5%) whilst progressing even more with logarithmic normalization (up to 4.8%). With respect to the NN, the results of the reduced parameters slightly decreased (up to less 4.3% for low aeration) whilst showing an increase with the best normalization procedure (up to 5.1% in low aeration) regarding the reduced results. Therefore, it could be found that regarding the critical conditions assessment for both statistical techniques the use of the normalization techniques provided better results than the ones obtained by the parameters set reduction.

4. CONCLUSIONS

The use of the complete set of raw data in the overall organism's recognition performance (species, genus, order, or sub-class identification) attained values of 85.8% (DA) and 84.5% (NN) for the non-stalked microorganisms and 51.9% (DA) and 49.9% (NN) for the stalked ones. Although the recognition performance for the non-stalked microorganisms

DOI: 10.1002/cem

could be regarded as quite fair, such was not the case for the stalked microorganisms performance, mainly due to the fact that some stalked species are morphologically speaking, hardly distinguishable.

Regarding the parameters reduction methodology, the use of a joint DTs and CA procedure resulted in 28 and 30% reductions in terms of the initial parameters set for the stalked and non-stalked microorganisms, respectively. Such reductions in the parameters set barely caused an effect in the results (differences below 3.2%), and may, therefore, be considered to form the basis of future analyses.

The logarithmic normalization previous to the application of DA technique caused slightly better results than the standard deviation normalization in both stalked and non-stalked microorganisms, however the differences below 3% can be considered of minor significance. Meanwhile, the normalization procedure for the NN results proved to present a slight improvement up to 2.1% for the standard deviation normalization in non-stalked microorganisms whereas for stalked ones the results have fallen. Therefore, although no advantages were found for the stalked microorganisms, regarding the non-stalked ones, the standard deviation normalization may be considered in future works. Hence, the best normalization procedures were found to be the logarithmic in the DA technique for both stalked and non-stalked microorganisms, whilst for the NN, the standard normalization for the non-stalked, and no normalization for the stalked were the most appropriate.

Regarding the identification of the main protozoa and metazoa groups (flagellate protozoa, ciliate protozoa, sarcodine protozoa, and metazoa) as well as the ciliated protozoa groups (carnivorous, crawling, free-swimming, and sessile) the results could be considered as quite fair. Indeed, the results for the protozoa and metazoa groups, protozoa ciliates, and sludge age assessment were rather good which is particularly important since ciliates are crucial for wastewater treatment plant diagnosis. Furthermore, for the assessment of the effluent quality, aeration, and nitrification, the results proved to be promising.

Moreover, considering the reduction in parameters set it was found that it resulted for the DA techniques in a non-significant decrease (up to 0.5%) in the overall protozoa and metazoa as well as ciliated protozoa groups identification, whereas for the NN motivated a small increase in the non-stalked overall recognition up to 2.6%.

Similarly, for the assessment of operational WWTP conditions, the parameters reduction for both multivariable statistical techniques did not affect significantly the overall results for the operational parameters survey when the DA is considered and a slight increase of up to 3.1% was observed for the NN. The use of the most favorable normalization (logarithmic) procedure led to a small increase in the aeration and sludge age assessment, up to 2.4% for the DA. For NN most favorable normalization study (no normalization for stalked and standard deviation for the non-stalked), no significant improvement in the results was found for the operational parameters assessment.

With respect to the critical conditions determination, the use of the reduced data set in DA allowed a minor improvement on the results progressing even more with logarithmic normalization. For NN, the results of the reduced parameters slightly decreased whilst showing an increase with the best normalization procedure.

Consequently, the use of the reduced parameters set has proven to be a suitable starting point for the both DA and NN methodologies, although for the DA analysis an initial logarithmic normalization step is advisable. For the NN analysis, a standard deviation normalization procedure could be considered for the non-stalked microorganisms regarding the operating parameters assessment.

As a general conclusion, image analysis coupled with a multivariate statistical technique such as DA proved to be a promising tool for assessing and monitoring protozoa and metazoa populations in a wastewater treatment plant. Furthermore, it was found that it is possible to reduce the size of the parameters data set used in previous analysis without significant loss of information, and advantages of pre-performing data normalization were also assessed for multiple cases.

Acknowledgements

The authors are grateful to the National Council of Scientific and Technological Development of Brazil (CNPq), the BI-EURAM III ALFA co-operation project (European Commission), and the POCI/AMB/57069/2004 project supported by the *Fundação para a Ciência e a Tecnologia* (Portugal). Data from Nancy plant made available by Prof. Maurício da Motta (UFPE, Recife, Brasil) is also acknowledged.

REFERENCES

1. Madoni P. A sludge biotic index (SBI) for the evaluation of the biological performance of activated-sludge plants based on the microfauna analysis. *Water Res.* 1994; **28**(1): 67–75.
2. Jahn TL, Bovee EC, Jahn FF. *How to Know the Protozoa*. Wm. C. Brown Company Publishers: Dubuque, 1999.
3. Fenchel T. *Ecology of Protozoa*. Springer-Verlag: Berlin, 1999.
4. Richard M. *Activated Sludge Microbiology*. The Water Pollution Control Federation: Alexandria, 1989.
5. Nicolau A, Dias N, Mota M, Lima N. Trends in the use of protozoa in the assessment of wastewater treatment. *Res. Microbiol.* 2001; **152**: 621–630.
6. Amaral AL, Baptiste C, Pons MN, Nicolau A, Lima N, Ferreira EC, Mota M, Vivier H. Semi-automated recognition of protozoa by image analysis. *Biotechnol. Tech.* 1999; **13**: 111–118.
7. da Motta M, Pons MN, Vivier H, Amaral AL, Ferreira EC, Roche N, Mota M. Study of protozoa population in wastewater treatment plants by image analysis. *Braz. J. Chem. Eng.* 2001; **18**: 103–111.
8. Golz C, Lange S. *Analysis of activated sludge using pattern recognition techniques. Proceedings of the 3rd International Specialized Conference on Micro-organisms in Activated Sludge and Biofilm Processes*, CD-ROM, Rome, 2001.
9. Einax JW, Zwanziger HW, Geiss S. *Chemometrics in Environmental Analysis*. VCH Verlagsgesellshaft: Weinheim, 1997.
10. Kasabov NK. *Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering*. The MIT Press: Cambridge, MA, 1996.
11. Amaral AL, da Motta M, Pons MN, Vivier H, Roche N, Mota M, Ferreira EC. Survey of protozoa and metazoa

DOI: 10.1002/cem

populations in wastewater treatment plants by image analysis and discriminant analysis. *Environmetrics* 2004; **15**: 381–390.
12. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man. Cybernetics.* 1979; **1**: 62–66.
13. Pons MN, Vivier H, Dodds J. Particle shape characterization using morphological descriptors. *Part. Syst. Charac.* 1997; **14**: 272–277.
14. Pons MN, Vivier H. Biomass quantification by image analysis. *Adv. Biochem. Eng. Biotechnol.* 1999; **66**: 133–184.
15. Obert M, Pfeifer P, Sernetz M. Microbial growth patterns described by fractal geometry. *J. Bacteriol.* 1990; **172**: 1180–1185.
16. Soddell JA, Sevior RJ. A comparison of methods for determining the fractal dimensions of colonies of filamentous bacteria. *Binary* 1994; **6**: 21–31.

DOI: 10.1002/cem