Analytica Chimica Acta 642 (2009) 94–101

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Correlation between sludge settling ability and image analysis information using partial least squares

D.P. Mesquita[a], O. Dias[a], A.M.A. Dias[a], A.L. Amaral[a,b], E.C. Ferreira[a,*]

[a] IBB-Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
[b] Instituto Superior de Engenharia de Coimbra, Instituto Politécnico de Coimbra, Rua Pedro Nunes, Quinta da Nora, 3030-199 Coimbra, Portugal

ARTICLE INFO

Article history:
Received 10 September 2008
Received in revised form 1 November 2008
Accepted 19 March 2009
Available online 24 March 2009

Keywords:
Activated sludge
Image analysis
Sludge volume index
Partial least squares

ABSTRACT

In the last years there has been an increase on the research of the activated sludge processes, and mainly on the solid–liquid separation stage, considered of critical importance, due to the different problems that may arise affecting the compaction and the settling of the sludge. Furthermore, image analysis procedures are, nowadays considered to be an adequate method to characterize both aggregated and filamentous bacteria, and increasingly used to monitor bulking events in pilot plants. As a result of that, in this work, image analysis routines were developed in *Matlab* environment, allowing the identification and characterization of microbial aggregates and protruding filaments. Moreover, the large amount of activated sludge data collected with the image analysis implementation can be subsequently treated by multivariate statistical procedures such as PLS. In the current work the implementation of image analysis and PLS techniques has shown to provide important information for better understanding the behavior of activated sludge processes, and to predict, at some extent, the sludge volume index. As a matter of fact, the obtained results allowed explaining the strong relationships between the sludge settling properties and the free filamentous bacteria contents, aggregates size and aggregates morphology, establishing relevant relationships between macroscopic and microscopic properties of the biological system.

1. Introduction

In activated sludge systems, an adequate balance between the different types of bacteria is necessary to ensure an efficient pollution removal, good sludge settling abilities and low suspended solid levels in the final effluent. After the oxidation of the organic matter in the aerated tank, the flocculated biomass is separated from the treated effluent by means of their settling ability in the settling tank. The settling phase is considered a critical stage of the process in which filamentous bulking and deflocculation processes are the most common problems, causing the decrease of the sludge settling ability and effluent quality deterioration [1]. Poor settling biomass is normally attained by an improper aggregate's formation and filamentous bacteria proliferation, resulting in lower clarifier efficiency. Usually, some malfunctions may occur within the activated sludge system such as pin-point flocs bulking, filamentous bulking, dispersed growth and zoogleal growth [2].

Image analysis procedures, based on microscopic observations, are nowadays considered to be a feasible method to characterize quantitatively aggregates and filamentous bacteria, and subsequently used to monitor bulking events in pilot and full plants [3,4], exceeding the initial manual quantification of filamentous bacteria proposed by Jenkins et al. [5]. Bulking can be caused by both filamentous and non-filamentous factors, affecting in different ways the sludge settling ability which can be detected by image analysis methodologies [4,6–12]. The combination of settling properties and the parameters obtained from image analysis may offer powerful information enabling immediate interventions on the biological system. In fact, the study developed by Sezgin [13] established that the sludge volume index (SVI) is strongly influenced by floc size and filamentous bacteria contents. Other authors [14–16] used automated image analysis to relate the microorganism's morphology in biological systems with the sludge settling properties. The settling ability can be subsequently related with the microscopic parameters using multivariable statistical technique, such as partial least squares (PLS) regression and principal component analysis (PCA) [4,6]. A close correlation between the filamentous bacteria per suspended solids ratio and the SVI was indeed achieved by Amaral and Ferreira [4] during filamentous bulking events.

Encouraged by the success of image analysis procedures over the last years in a broad range of different areas, the present work uses an automated image analysis method to characterize the activated sludge structure, focusing on the prediction ability of sludge settling properties, in both good settling and filamentous bulking periods. In this sense, the collected images were treated in order to characterize the aggregated and filamentous bacteria, thus originating

* Corresponding author. Tel.: +351 253 604 407; fax: +351 253 678 986.
E-mail address: ecferreira@deb.uminho.pt (E.C. Ferreira).

0003-2670/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2009.03.023

the different morphological parameters, further used as independent **X** variables in the PLS regression for the sludge volume index (SVI) model determination.

2. Materials and methods

2.1. Sampling survey

The activated sludge samples analyzed in this work were collected during several months from the aeration basins of eight municipal wastewater treatment plants (WWTP) treating domestic effluents located in the North of Portugal (Frossos1, Frossos2, Cunha, Sobreposta, Tadim, Pousa, Ucha, Oliveira). A total of 142 samples (representing different days in different WWTP) were collected (17 from Frossos1, 33 from Frossos2, 14 from Cunha, 23 from Sobreposta, 21 from Tadim, 11 from Pousa, 12 from Ucha and 11 from Oliveira) for a period between 4 and 7 months and time spans varying from 3 until 7 days between each sample. From each sample the total suspend solids (TSS) was measured by weight [17] and further used to calculate the sludge volume index (SVI) in a 10 L cylindrical column (15 cm diameter) for 30 min. This physical property was determined as follows:

$$\mathrm{SVI} = \frac{H_{30}}{H_0 \times \mathrm{TSS}} \quad (1)$$

where H_0 and H_{30} are the height in the time 0 and 30 min, respectively. The SVI is highly used for sludge settling ability evaluation, varying inversely with the sludge ability to settle, making it an interesting parameter within the system efficiency assessment.

Microscopic observations were performed for all samples in order to determine the morphological parameters of microbial aggregates and filament contents using digital image analysis. The maximum period of time between sample collection and image acquisition did not exceed 2 h with not more than 30 min without aeration.

2.2. Image acquisition

For image acquisition, a volume of 25 μL from each sample was distributed on a slide and covered with a 20 mm × 20 mm cover slip for visualization and image acquisition. The volume deposition was performed by means of a recalibrated micropipette with a sectioned tip at the end to a diameter allowing the passage of even the larger aggregates. During the survey, three different slides (replicates) were screened in order to minimize sampling errors, and images were acquired for 22 different positions in the upper, middle and lower parts of the slide. Therefore, a total around 200 images per sample (22 × 3 positions × 3 replicates) were acquired in bright field microscopy to obtain representative information of the sludge. All the images were acquired in a *Leitz Laborlux S* optic microscope (*Leitz, Wetzlar*), with 100× magnification, coupled to a *Zeiss AxioCam HRc* (*Zeiss, Oberkochen*) camera. The image acquisition was performed in 1300 × 1030 pixels and 8-bit format through the *AxioVision 3.1* (*Zeiss, Oberkochen*) software. This acquisition methodology was performed for Frossos2, Cunha, Sobreposta, Tadim, Pousa, Ucha and Oliveira WWTP.

A previous survey studied by Amaral and Ferreira [4] was also included in this work representing an earlier period of filamentous bulking conditions in one of the studied WWTP (described as Frossos1 samples). The image acquisition of the Frossos1 survey relied on 20 images per sample (three slides were also used with the purpose described above) acquired in bright field microscopy in a *SZ 4045TR-CTV Olympus* stereo microscope (*Olympus, Tokyo*), with 40× magnification, coupled to a *Sony CCD AVC–D5CE* (*Sony, Tokyo*) grey scale video camera. The image acquisition was performed in 768 × 576 pixels and 8-bit format by a *Data Translation DT 3155* (*Data Translation, Marlboro*) frame grabber using the commercial software package *Image Pro Plus* (*Media Cybernetics, Silver Spring*). In order to compare both acquisition methodologies, calibration from pixels to the metric unit dimensions was performed by means of a micrometer slide.

2.3. Image processing

The image processing and analysis program for aggregated and filamentous bacteria characterization was developed in *Matlab* 7.3 (*The Mathworks, Inc., Natick*) language, adapting a previous version developed by Amaral [18]. Primarily, the image processing step established the binary images from the aggregated biomass and the protruding and freely dispersed filamentous bacteria and thereafter, morphological parameters were determined. In total, and during the current work, a universe around 400,000 aggregates, from the overall 142 samples dataset of the 8 WWTPs, was acquired and individually characterized.

Fig. 1 shows a schematic representation of the main steps of the program, comprising the image pre-treatment, segmentation, and debris elimination whereas the image analysis program was oriented to the aggregated and filamentous bacteria characterization and contents determination.

2.4. Image analysis parameters

Following the image processing step, the recognized aggregated and filamentous bacteria from the collected images, were treated in order to individually characterize each aggregate and filamentous bacteria, in terms of the most relevant morphological parameters. Therefore, and based on the previous study of Amaral and Ferreira [4], 36 morphological parameters were determined (in addition to the 2 sludge physical properties, TSS and SVI), either directly from the image analysis program, either in association with the sludge physical properties. The overall dataset of 38 parameters is presented in Table 1.

Total aggregates number per sludge volume (Nb/Vol), total aggregates area per sludge volume (TA/Vol), aggregates individual area (Area) and total filaments length per sludge volume (TL/Vol), described in [18], were determined for all the samples collected alongside the filaments individual length (FL), aggregates individual length (L), aggregates individual perimeter (P) and aggregates individual equivalent diameter (D_{eq}) described next.

The filaments individual length (FL) was determined according to [19], with N_{Thn} as the pixel sum of each thinned filament, N_{int} as the number of filaments intersections and factor 1.1222 used to average the different measuring angles within the image:

$$\mathrm{FL} = (N_{\mathrm{Thn}} + N_{\mathrm{int}})1.1222 \times F_{\mathrm{Cal}} \quad (2)$$

The aggregates individual equivalent diameter (D_{eq}) was calculated based on the area determination with F_{Cal} as the calibration factor (μm pixel^{-1}) [20,21]:

$$D_{\mathrm{eq}} = 2F_{\mathrm{Cal}}\sqrt{\frac{\mathrm{Area}}{\pi}} \quad (3)$$

The aggregates individual perimeter was calculated by [19] with N_{Per} as the pixel sum of the objects and factor 1.1222 used to average the different measuring angles within the image:

$$P = N_{\mathrm{Per}} \times 1.1222 \times F_{\mathrm{Cal}} \quad (4)$$

The aggregates individual length (L) was given by the maximum Feret Diameter, which is the maximum distance between two parallel tangents touching opposite borders of the object, converted to metric units [20,21].

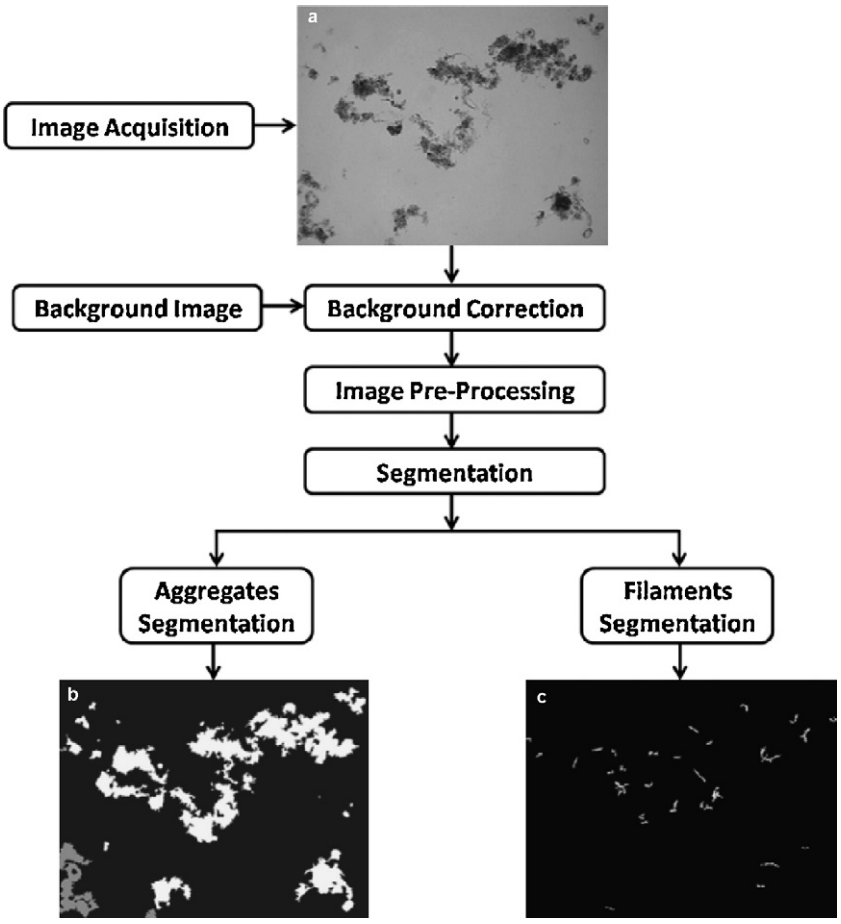

Fig. 1. Schematic representation of the image analysis procedures.

The morphological descriptors convexity (Conv), solidity (Solid), roundness (Round), and eccentricity (Ecc) were also determined for each individual aggregate, by the image analysis methodology, as follows.

Eccentricity was calculated using the area and the second order moments (M_2) of each individual aggregate as [20]:

$$\text{Ecc} = \frac{(4\pi)^2(M_{2X} - M_{2Y})^2 + 4M_{2XY}^2}{\text{Area}^2} \quad (5)$$

Convexity was determined as the ratio between the convex perimeter (P_{Conv}) and the perimeter (P) of each individual aggregate by [20]:

$$\text{Conv} = \frac{P_{\text{Conv}}}{P} \quad (6)$$

Roundness was calculated using the area and the convex perimeter of each individual aggregate by [20]:

$$\text{Round} = \frac{4\pi\text{Area}}{P_{\text{Conv}}^2} \quad (7)$$

Solidity was determined as the ratio between each individual aggregate area and convex envelope area ($\text{Area}_{\text{Conv}}$) [21].

$$\text{Solid} = \frac{\text{Area}}{\text{Area}_{\text{Conv}}} \quad (8)$$

Furthermore, the aggregates morphological characterization was subsequently divided in three classes in order to differentiate between small low settling structures (contained within the aggregates below 0.025 mm in equivalent diameter), intermediate good settling structures (between 0.025 and 0.25 mm in equivalent diameter), and large connected low settling structures (contained within the aggregates above 0.25 mm in equivalent diameter). For each of these aggregates class (small, intermediate and large) the number of aggregates per volume was determined as the total number of the aggregates belonging to that class present in a given activated sludge volume, as well as the aggregates area percentage as the ratio between the total area of the aggregates belonging to that class and the total area of all aggregates, Those parameters were determined in order to account for the contents significance of the each aggregates class, both in terms of their number and area.

Finally, total filaments length per sludge volume (TL/Vol), filaments average length per aggregates average area ratio ($\text{FL}_{\text{avg}}/\text{Area}_{\text{avg}}$), filaments average length per aggregates average equivalent diameter ratio ($\text{FL}_{\text{avg}}/D_{\text{eq}_{\text{avg}}}$) and the total filaments length per total aggregated area ratio (TL/TA) were determined alongside the total filaments length per volatile suspended solids (TL/TSS) ratio characterizing the filaments dynamics.

For clarity purposes, the dataset was classified, according to the physical and morphological meaning of each parameter, in five main descriptor groups: free filamentous bacteria contents; free filamentous bacteria characterization; aggregates content; aggregates size and aggregates morphology. Furthermore, for the aggregates content, aggregates size, and aggregates morphology, a more detailed analysis was performed, including data of all aggregates, labeled as overall (ovr), intermediate (int) and large aggregates (lrg). For the determination of these values, the average (mean value) of, respectively, all (ovr), intermediate (int) and large (lrg) individual aggregate parameters, was determined. Due to the poor pixel representation of the small aggregates within

Table 1
Parameters collected from the eight wastewater treatment plants.

Parameter	Symbol	Description
Predictor (**X**)	*Free filamentous bacteria contents*	
	TL/Vol	Total filaments length per volume
	TL/TSS	Total filaments length per total suspended solids
	TL/TA	Total filaments length per total aggregates area
	Free filamentous bacteria characterization	
	$FL_{avg}/Area_{avg}$	Filaments average length per aggregates average area
	FL_{avg}	Filaments average length
	$FL_{avg}/D_{eq_{avg}}$	Filaments average length per aggregates average equivalent diameter
	Aggregates content	
	TA/Vol	Total aggregates area per volume
	Nb_{tot}/Vol	Total number of aggregates per volume
	Nb_{sml}/Vol	Number of small aggregates per volume
	Nb_{int}/Vol	Number of intermediate aggregates per volume
	Nb_{lrg}/Vol	Number of larger aggregates per volume
	TSS	Total suspended solids
	Aggregates size	
	Overall	
	$Area_{ovr}$	Aggregates average area
	% $Area_{sml}$	Area percentage of small aggregates
	% $Area_{int}$	Area percentage of intermediate aggregates
	% $Area_{lrg}$	Area percentage of larger aggregates
	$D_{eq_{ovr}}$	Aggregates average equivalent diameter
	P_{ovr}	Aggregates average perimeter
	L_{ovr}	Aggregates average length
	Intermediate aggregates	
	$D_{eq_{int}}$	Equivalent diameter of intermediate aggregates (mean value)
	P_{int}	Perimeter of intermediate aggregates (mean value)
	L_{int}	Length of intermediate aggregates (mean value)
	Large aggregates	
	$D_{eq_{lrg}}$	Equivalent diameter of larger aggregates (mean value)
	P_{lrg}	Perimeter of larger aggregates (mean value)
	L_{lrg}	Length of larger aggregates (mean value)
	Aggregates morphology	
	Overall	
	$Conv_{ovr}$	Aggregates average Convexity
	$Round_{ovr}$	Aggregates average Roundness
	$Solid_{ovr}$	Aggregates average Solidity
	Ecc_{ovr}	Aggregates average Eccentricity
	Intermediate aggregates	
	$Conv_{int}$	Convexity of intermediate aggregates (mean value)
	$Round_{int}$	Roundness of intermediate aggregates (mean value)
	$Solid_{int}$	Solidity of intermediate aggregates (mean value)
	Ecc_{int}	Eccentricity of intermediate aggregates (mean value)
	Large aggregates	
	$Conv_{lrg}$	Convexity of large aggregates (mean value)
	$Round_{lrg}$	Roundness of large aggregates (mean value)
	$Solid_{lrg}$	Solidity of large aggregates (mean value)
	Ecc_{lrg}	Eccentricity of large aggregates (mean value)
Response (**Y**)	SVI	Sludge volume index

the collected images, due to their small size, the determination of their morphological properties is quite prone to large errors, and therefore, their morphological parameters were not included in this study. In order to simplify the identification, Table 1 shows all the parameters description taken from the program including those associated with physical properties.

2.5. Data reduction and partial least squares

A cross-correlation analysis between the collected data was then performed in order to reduce the dataset, leading to the exclusion of one variable (physical or morphological parameter) for each pair presenting a correlation factor above 90%.

The reduced dataset was further treated by partial least squares (PLS) regression, an iterative algorithm that extracts linear combinations of the essential features of the original **X** data while modeling the **Y** data dependence on the work set, in order to perform a multivariate calibration of the SVI variable (**Y**) from the remaining dataset (**X** variables) [22,23]. PLS has been shown to be an efficient approach in monitoring complex processes since the high dimensional strongly cross-correlated data can be reduced to a much smaller and interpretable set of latent variables [22,24]. PLS reduces the dimension of the predictor variables by extracting factors or latent variables that are correlated with **Y** while capturing a large amount of the variations in **X**. This means that PLS maximizes the covariance between matrices **X** and **Y**. In PLS, the scaled matrices **X** and **Y** are decomposed into score (**t** and **u**) and loading (**p** and **q**) vectors, and residual error matrices (**E** and **F**):

$$\mathbf{X} = \sum_{i=1}^{a} \mathbf{t}_i \mathbf{p}_i^T + \mathbf{E} \tag{9}$$

$$\mathbf{Y} = \sum_{i=1}^{a} \mathbf{u}_i \mathbf{q}_i^T + \mathbf{F} \tag{10}$$

where a is the number of latent variables. In an inner relation the score vector **t** is linearly regressed against the score vector **u**.

$$\mathbf{u}_i = b_i \mathbf{t}_i + h_i \tag{11}$$

where b is regression coefficient which is determined by minimizing the residual h.

It is crucial to determine the optimal number of latent variables (component) and cross-validation (CV) is a practical and reliable way to test the predictive significance of each PLS regression. In order to do so, part of the training dataset is kept out of the model development, predicted by the model and finally compared with the actual values (cross-validation). The prediction error sum of squares (PRESS) is the sum of the squared differences between the observed and predicted values for the data kept out of the model fitting. This procedure is repeated until every sample has been kept out once and only once. Therefore, the final PRESS has contributions from all data [22,23].

To perform the PLS analysis from the data set, *SIMCA 8.0* (*Umetrics, Umeå*) software package was used. This software first mean centers the data and scales it by the use of the standard deviation. A cross-validation method is used in which 7 subgroups are sequentially removed from the model development in each new PLS latent variable (component) determination. Furthermore, for every dimension, *SIMCA* computes the overall PRESS/SS, where SS is the residual sum of squares of the previous dimension. A component is then considered significant if PRESS/SS is statistically smaller than 1.0. *SIMCA* also displays the Q^2 value which is the fraction of the total variation of the **X**'s that can be predicted by a component, as estimated by cross-validation, and is computed as (1.0-PRESS/SS). Furthermore, *SIMCA* also determines the variable importance in the projection (VIP), which represents the influence of each variable (k) of the data matrix (**X**) on the results matrix (**Y**), so that the variables with a VIP larger than 1 have an above average influence on the result and are, therefore, the most relevant for explaining **Y**. *SIMCA* computes the VIP as the sum over all model dimensions (PLS components) of the variable influence. A more detailed description about this method can be found in [25].

In the present study, the predictor matrix **X** consisted of 37 variables and the response matrix **Y** of the key variable SVI (Table 1) in a total of 142 input samples in total. The sludge volume index was selected as **Y** variable since it is a key indicator of the sludge settling

Table 2
Variables presenting correlation factors above 0.9 and respective VIP values for the PLS regression with 3 components.

Variable kept	(VIP)	Variable excluded	(VIP)	Correlation
TL/TSS	2.111	TL/TA	1.995	0.9265
$FL_{avg}/D_{eq_{avg}}$	0.838	$FL_{avg}/Area_{avg}$	0.725	0.9041
% $Area_{sml}$	1.113	$FL_{avg}/D_{eq_{avg}}$	0.838	0.9075
Nb_{sml}/Vol	0.860	Nb_{tot}/Vol	0.567	0.9244
$D_{eq_{ovr}}$	1.269	$Area_{ovr}$	0.916	0.9356
L_{ovr}	1.276	L_{lrg}	1.265	0.9305
P_{ovr}	1.080	P_{lrg}	1.060	0.9823
L_{int}	0.166	$D_{eq_{int}}$	0.161	0.9238
P_{int}	0.240	L_{int}	0.166	0.9318
Ecc_{ovr}	1.009	Ecc_{lrg}	0.844	0.9247
$Round_{ovr}$	0.986	$Solid_{ovr}$	0.854	0.9715
$Round_{lrg}$	1.080	$Round_{ovr}$	0.986	0.9474
$Round_{lrg}$	1.080	$Solid_{ovr}$	0.854	0.9039
$Solid_{ovr}$	0.854	$Solid_{lrg}$	0.635	0.9380

ability. For the current PLS application, a random 95 samples were used for training (2/3 of the dataset), and the remaining 47 samples (1/3 of the dataset) for the validation of the developed models.

3. Results and discussion

The cross-correlation analysis on the full dataset (5396 data points from 142 samples × 38 variables, including 1 **Y** variable) allowed to exclude 13 **X** variables (TL/TA, $FL_{avg}/Area_{avg}$, $FL_{avg}/D_{eq_{avg}}$, Nb_{tot}/Vol, $Area_{ovr}$, $D_{eq_{int}}$, L_{int}, P_{lrg}, L_{lrg}, $Round_{ovr}$, $Solid_{ovr}$, $Solid_{lrg}$ and Ecc_{lrg}) due to the existence of a correlation factor above 0.9 with other variables. The exclusion of one variable, for each pair presenting a correlation factor above 0.9, was determined after a full 38 variables dataset PLS regression was performed, allowing to establish the most important variable (higher VIP), of the pair, for the SVI prediction. The results of the PLS regression variables importance (VIP) for 3 components (best fit according to Q^2 values), and the correlation factors, for each studied pair of variables is presented in Table 2. The remaining variables were automatically used for the further approach leading to a reduced dataset of 25 variables (24 **X** variables and 1 **Y** variable) used in the subsequent PLS analysis.

Subsequently a second PLS regression was applied using the reduced 25 variables dataset (24 **X** variables), to describe the SVI, and seemingly evaluated for a number of components (latent variables) ranging up to 22 (until a negligible correlation increase between **X** and **Y** was found). According to the results, presented in Table 3, a total of 2 components (latent variables) allowed to obtain the best SVI model, given the fact that only the two first components presented Q^2 values higher than the 0.097 limit (cross-validation threshold for the current 95 samples training dataset PLS analysis).

However, it was found that the 2 components PLS regression presented quite low cumulative fractions on the variation of the **Y**'s variables (0.7257 for cumulative R^2Y), for a cumulative Q^2 of 0.6716. Furthermore, the results obtained for the regression analysis between the predicted and observed SVI values (Table 3), were shown to present a mediocre prediction ability with a correlation coefficient of 0.9155 (R^2 of 0.8381) for the validation dataset, and even lower for the training dataset with a correlation coefficient of 0.8519 (R^2 of 0.7257). Also, when a regression analysis between the predicted and observed SVI values from both the validation and the training datasets was performed (training + validation results in Table 3) a mediocre prediction ability was found with a correlation coefficient of 0.8720 (R^2 of 0.7604). The inclusion of further components on the model was studied, taking into account overfitting problems that may arise by the inclusion of too many components. Thus, the number of components was chosen as the best compromise of high correlation coefficient and slope close to one, and always taking the validation set as reference. Analyzing Table 3, it was found that the 6 components PLS regression presented the best set of conditions regarding the combination of high correlation coefficient (0.9138) and slope value close to one (0.8751). Fig. 2 presents the regression analysis between the predicted and observed SVI values for the 25 variables dataset (24 **X** variables) PLS regression with 6 components, where the attained aggregated dataset correlation coefficient was 0.9289 (R^2 of 0.8628) for a regression slope of 0.8522.

Subsequently, a selection was carried out, regarding the variables presenting variable importance (VIP) values higher than 1, in order to perform a third PLS study, resulting in a total of 10 variables for both 2 and 6 components PLS regression, as presented in Table 4. The results showed that the TL/TSS was the most important variable (larger VIP), corroborating the findings of Amaral and Ferreira [4], followed by the TL/Vol, both representing the free filamentous bac-

Table 3
Values of the cumulative R^2X, R^2Y and Q^2 values for the model and slope and regression factors for the training, validation and training + validation datasets, for the 25 variables dataset PLS regression (slope values are equal to the R^2 values in the model built on the training set).

Component	Model (training set)				Validation set		Training + validation sets	
	Slope and R^2	R^2X	R^2Y	Q^2	Slope	R^2	Slope	R^2
1	0.6506	0.2780	0.6506	0.6240	0.7819	0.7910	0.6920	0.6941
2	0.7257	0.1468	0.0751	0.1266	0.8410	0.8381	0.7605	0.7604
3	0.7722	0.1393	0.0465	0.0895	0.8567	0.8450	0.7979	0.7949
4	0.8055	0.0882	0.0333	−0.0421	0.8589	0.9043	0.8225	0.8355
5	0.8180	0.1572	0.0125	0.0295	0.8566	0.9166	0.8303	0.8476
6	0.8406	0.0307	0.0226	−0.1613	0.8751	0.9138	0.8522	0.8628
7	0.8526	0.0482	0.0120	−0.0251	0.8657	0.9240	0.8567	0.8738
8	0.8619	0.0238	0.0094	−0.0902	0.8745	0.9107	0.8650	0.8765
9	0.8651	0.0194	0.0032	−0.2013	0.8540	0.8936	0.8611	0.8735
10	0.8696	0.0081	0.0045	−0.2515	0.8752	0.9014	0.8709	0.8793
11	0.8726	0.0106	0.0030	−0.2433	0.8797	0.8895	0.8741	0.8778
12	0.8749	0.0102	0.0023	−0.1924	0.8951	0.8843	0.8812	0.8782
13	0.8758	0.0102	0.0009	−0.1354	0.9022	0.8825	0.8839	0.8782
14	0.8762	0.0060	0.0005	−0.1246	0.8955	0.8773	0.8822	0.8769
15	0.8765	0.0046	0.0003	−0.1131	0.8961	0.8760	0.8824	0.8767
16	0.8768	0.0043	0.0002	−0.1060	0.8992	0.8735	0.8834	0.8760
17	0.8769	0.0035	0.0001	−0.0978	0.9013	0.8761	0.8842	0.8769
18	0.8770	0.0035	8.6e-05	−0.0618	0.9054	0.8780	0.8856	0.8775
19	0.8770	0.0032	3.7e-05	−0.0617	0.9060	0.8771	0.8858	0.8773
20	0.8770	0.0026	7.8e-06	−0.0643	0.9063	0.8780	0.8859	0.8775
21	0.8770	0.0010	2.6e-06	−0.0315	0.9072	0.8777	0.8861	0.8774
22	0.8770	0.0005	5.1e-08	−0.0248	0.9073	0.8776	0.8861	0.8774

Table 4
Variables presenting a VIP larger than 1, for the 25 variables dataset PLS regression with 2 and 6 components.

Descriptor group	Variable	VIP 2 components	VIP 6 components
Free filamentous bacteria contents	TL/TSS	2.128	2.008
	TL/Vol	1.666	1.641
Aggregates size	L_{ovr}	1.293	1.233
	$D_{eq_{ovr}}$	1.270	1.226
	% $Area_{sml}$	1.106	1.084
	P_{ovr}	1.065	1.016
Aggregates morphology	$Solid_{int}$	1.402	1.410
	$Conv_{int}$	1.385	1.311
	$Round_{lrg}$	1.077	1.035
	Ecc_{ovr}	1.023	1.076

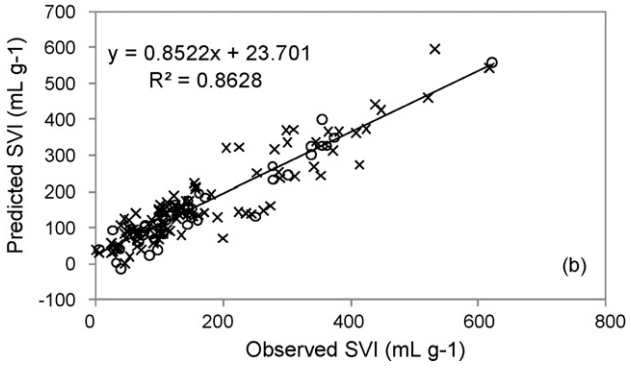

Fig. 2. Relationship between the predicted and observed SVI for the 25 variables dataset PLS regression with 6 components. Cross marks represent the training dataset and the circles the validation dataset.

teria contents, thus clearly stating the importance of this descriptor group. It was also found that 8 other **X** variables presented a VIP higher than 1 including the intermediate aggregates convexity and solidity, the overall eccentricity and the large aggregates roundness for the aggregates morphology descriptor group, the overall equivalent diameter, perimeter and length, and the small aggregates area % for the aggregates size descriptor group. Furthermore, from these results it could be established that both the free filamentous aggregates characterization and the aggregates content seemed to play no significant role in predicting the SVI, as no variable from these descriptor groups had a VIP larger than 1 for a 2 and 6 components PLS regressions.

Subsequently a third PLS regression was applied using the 11 variables dataset (10 **X** variables and 1 **Y** variable), to describe the SVI, and seemingly evaluated for a number of components (latent variables) ranging up to 10 (number of **X** variables). According to the results, presented in Table 5, a total of 2 components (latent variables) allowed to obtain the best SVI model, given the fact that only the two first components presented Q^2 values higher than the 0.097 Q^2 limit (cross-validation threshold for the current 95 samples training dataset PLS analysis).

It was found that the 2 components PLS repression presented quite low cumulative fractions on the variation of the **Y**'s variables (0.7686 for cumulative R^2Y), for a cumulative Q^2 of 0.7385. Furthermore, the results obtained for the regression analysis between the predicted and observed SVI values (Table 5), were shown to present a mediocre prediction ability with a correlation coefficient of 0.9112 (R^2 of 0.8302) for the validation dataset, and even lower for the training dataset with a correlation coefficient of 0.8767 (R^2 of 0.7686). Also, when a regression analysis between the predicted and observed SVI values from both the validation and the training datasets was performed (training + validation results in Table 5) a mediocre prediction ability was found with a correlation coefficient of 0.8963 (R^2 of 0.8034). Following the same strategy of the full model, the number of components was increased to 5 regarding the combination of high correlation coefficient (0.8874) and slope value close to one. Fig. 3 presents the regression analysis between the predicted and observed SVI values for the 11 variables dataset (10 **X** variables) PLS regression with 5 components, where the attained aggregated dataset correlation coefficient was 0.9186 (R^2 of 0.8439) for a regression slope of 0.8332. With these results it is clear that the decrease in dimensionality from the full data set model to models with 24 and 10 variables does not significantly worsen the SVI prediction ability in terms of correlation coefficient or regression slope in the representation of Y_{pred} vs $Y_{measured}$.

However it should be kept in mind that the predicted SVI values are still far from the precision of the original measurement. As a matter of fact, the error of the SVI calculation was determined as 5.7%, which for the 166 mL g^{-1} average SVI in this study signifies that an average error of 9.462 mL g is expected for the observed SVI values. Comparing to the root mean error of prediction (RMSEP) obtained by the 37 **X** variables dataset (49.705 mL g^{-1} for 2 components and 36.576 mL g^{-1} for 6 com-

Table 5
Values of the cumulative R^2X, R^2Y and Q^2 values for the model and slope and regression factors for the training, validation and training + validation dataset, for the 11 variables dataset PLS regression (slope values are equal to the R^2 values in the model built on the training set).

Component	Model (training set)				Validation set		Training + validation sets	
	Slope and R^2	R^2X	R^2Y	Q^2	Slope	R^2	Slope	R^2
1	0.6565	0.5364	0.6565	0.6401	0.7516	0.8061	0.6873	0.7018
2	0.7686	0.1111	0.1121	0.2732	0.8302	0.8837	0.7875	0.8034
3	0.8046	0.1269	0.0360	0.0828	0.8278	0.8803	0.8122	0.8275
4	0.8152	0.0793	0.0106	−0.0292	0.8343	0.8917	0.8221	0.8381
5	0.8247	0.0443	0.0095	−0.0505	0.8518	0.8874	0.8332	0.8439
6	0.8291	0.0391	0.0045	−0.0738	0.8399	0.8837	0.8322	0.8456
7	0.8308	0.0151	0.0017	−0.1423	0.8455	0.8852	0.8349	0.8472
8	0.8311	0.0224	0.0003	−0.1220	0.8472	0.8840	0.8355	0.8471
9	0.8311	0.0122	4.0e-05	−0.0810	0.8477	0.8854	0.8357	0.8475
10	0.8312	0.0132	7.3e-07	−0.0546	0.8475	0.8853	0.8357	0.8475

Table 6
Variables importance values for the 11 variables dataset PLS regression for 2 and 5 components.

Descriptor group	Variable	VIP 2 components	VIP 5 components
Free filamentous bacteria contents	TL/TSS	1.568	1.516
	TL/Vol	1.161	1.165
Aggregates size	L_{ovr}	0.949	0.919
	$D_{eq_{ovr}}$	0.897	0.912
	% $Area_{sml}$	0.760	0.777
	P_{ovr}	0.851	0.848
Aggregates morphology	$Solid_{int}$	0.981	1.048
	$Conv_{int}$	0.996	1.011
	$Round_{lrg}$	0.827	0.820
	Ecc_{ovr}	0.740	0.742

ponents) and by the 10 **X** variables dataset (42.614 mL g^{-1} for 2 components and 41.598 mL g^{-1} for 5 components), it seems clear that the predicted SVI is still somewhat far from the precision of the observed SVI. However, it should be stressed that the PLS prediction, apart from providing a gross SVI estimation as a function of the aggregate parameters, above all, helps to identify the most relevant parameters linked to the increase or decrease of the SVI.

Finally, the most important variables regarding SVI prediction (higher VIP values) were sought, resulting in a total of 2 variables for the 2 components PLS regression, and 4 variables for the 5 components PLS regression, as presented in Table 6. The results showed that the TL/TSS was the most important variable (larger VIP), once again corroborating the findings of Amaral and Ferreira [4], followed by the TL/Vol, both representing the free filamentous bacteria contents, thus clearly stating the importance of this descriptor group. It was also found that 2 other **X** variables presented a VIP higher than 1 in the 5 components PLS regression, namely the intermediate aggregates convexity and solidity both regarding the aggregates morphology descriptor group. In addition, the regression coefficients of 0.6083, 0.1561, −0.4953 and 0.2240, for the 5 components PLS regression, regarding respectively the TL/TSS, TL/Vol, intermediate aggregates solidity and intermediate aggregates convexity, allowed to establish the relationships between these variables and the SVI. In that sense, it could be established that an increase on the filaments content and filaments content per suspended solids, as well as on the intermediate aggregate borders roughness (convexity) seems to lead to the increase of the SVI and, therefore of low settling abilities. Inversely, the increase on the intermediate aggregates solidity seems to lead to the decrease of the sludge volume index. Moreover, the present relationships could be expected for filamentous bulking phenomena, characterized by high filaments to aggregate ratios.

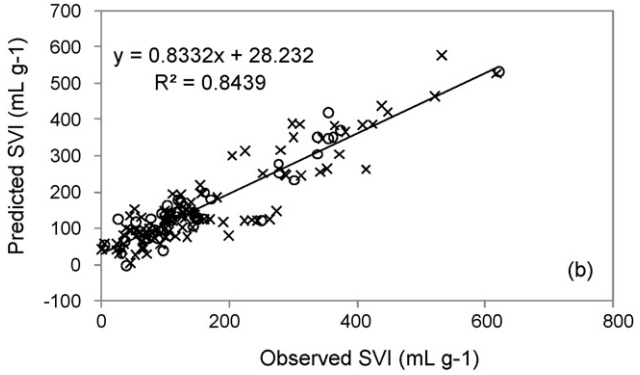

Fig. 3. Relationship between the predicted and observed SVI for the 11 variables dataset PLS regression with 5 components. Cross marks represent the training dataset and the circles the validation dataset.

4. Conclusions

Relating to the work of Amaral and Ferreira [4], in the course of this survey, it was possible to study a wider range of SVI data, comprising good and poor settling ability properties of the sludge, whereas in the earlier study the SVI values comprised solely values larger than 250 mL g^{-1}. In this way, and given the attained correlation coefficients and established relationships, it seems reasonable to infer that the PLS regressions performed during this study showed some promising results.

Moreover, the obtained results allowed explaining the strong relationships between the sludge settling properties, as described by the SVI, and some filamentous bacteria and aggregated biomass morphological descriptor groups such as the free filamentous bacteria contents, aggregates size and aggregates morphology, establishing relevant relationships between macroscopic and microscopic properties of the biological system. Furthermore, the performed parameters reduction study, allowed the establishment of a 10 independent variables procedure for SVI estimation, after the identification of the most important variables regarding SVI prediction, and the demonstration of the utmost importance of the free filamentous bacteria contents descriptor group on the SVI prediction ability.

Acknowledgements

The authors acknowledge the financial support to Daniela Mesquita and Oscar Dias through the grant SFRH/BD/32329/2006 and the project POCI/AMB/57069/2004, respectively, provided by Fundação para a Ciência e Tecnologia (Portugal). The authors express their gratitude to AGERE (Empresa de Águas, Efluentes e Resíduos de Braga, Portugal—EM) and Rui Gonçalves for their cooperation.

References

[1] X. Li, Y. Yuan, Water Res. 36 (2002) 3110.
[2] G. Bitton (Ed.), Wastewater Microbiology, Wiley/Liss, New York, 1994.
[3] M. da Motta, M.N. Pons, N. Roche, Water Sci. Technol. 43 (7) (2001) 91.
[4] A.L. Amaral, E.C. Ferreira, Anal. Chim. Acta 544 (2005) 246.
[5] D. Jenkins, M.G. Richard, G.T. Daigger, Manual on the Causes and Control of Activated Sludge Bulking and Foaming, WRC, Pretoria and USEPA, Cincinnati, 1986.
[6] R. Jenné, E.N. Banadda, J. Deurinck, I.Y. Smets, A.H. Geeraerd, J.F. Van Impe, Water Sci. Technol. 54 (1) (2006) 167.
[7] R. Jenné, E.N. Banadda, I.Y. Smets, J.F. Van Impe, Water Sci. Technol. 50 (2004) 281.
[8] R. Jenné, E.N. Banadda, I.Y. Smets, J. Deurinck, J.F. Van Impe, Microsc. Microanal. 13 (2007) 36.
[9] M. da Motta, A.L. Amaral, M. Casellas, M.N. Pons, C. Dagot, N. Roche, E.C. Ferreira, H. Vivier, IFAC Comput. Appl. Biotechnol. (2001) 427.
[10] M. da Motta, M.N. Pons, N.N. Roche, H. Vivier, Biochem. Eng. J. 9 (2001) 165.
[11] J.C. Palm, D. Jenkins, D.S. Parker, J. Water Pollut. Control Fed. 52 (10) (1980) 2484.
[12] C. Dagot, M.N. Pons, M. Casellas, G. Guibaud, P. Dollet, M. Baudu, Water Sci. Technol. 43 (3) (2001) 27.

[13] M. Sezgin, Water Sci. Technol. 16 (1982) 83.
[14] S. Matsui, R. Yamamoto, Water Sci. Technol. 16 (1984) 69.
[15] J.J. Ganczarczyk, Water Sci. Technol. 30 (1994) 87.
[16] K. Grijspeerdt, W. Verstraete, Water Res. 31 (1997) 1126.
[17] APHA, AWWA, WPCF, Standard Methods for the Examination of Water and Wastewater, American Public Health Association, Washington, DC, 1989.
[18] A.L. Amaral. Image analysis in biotechnological processes: applications to wastewater treatment, Ph.D. Thesis, Braga, Portugal, 2003, http://hdl.handle.net/1822/4506.
[19] A.E. Walsby, A. Avery, J. Microbiol. Methods 26 (1996) 11.
[20] C.A. Glasbey, G.W. Horgan, Image Analysis for the Biological Sciences, Wiley, Chichester, 1995.
[21] C.R. Russ, The Image Processing Handbook, CRC Press, Boca Raton, FL, 1995.
[22] P. Teppola, S.-P. Mujunen, P. Minkkinen, Chem. Intell. Lab. Syst. 38 (1997) 197.
[23] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109.
[24] P. Aarnio, P. Minkkinen, Anal. Chim. Acta 191 (1986) 457–460.
[25] A.B. Umetri, User's Guide to SIMCA, 1998, CD-ROM.