Predicting effects of Toxic Events to Anaerobic Granular Sludge with Quantitative Image Analysis and Principal Component Analysis

J.C. Costa, M.M. Alves, E.C. Ferreira

IBB – Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal. (E-mail: *ecferreira@deb.uminho.pt*)

Abstract
Detergents and solvents are included in the list of compounds that can be inhibitory or toxic to anaerobic digestion processes. Industrial cleaning stages/processes produce vast amounts of contaminated wastewater. In order to optimize the control of these wastewaters it is important to know and predict the effects on the activity and physical properties of anaerobic aggregates in an early stage. Datasets gathering morphological, physiological and reactor performance information were created from three toxic shock loads (SL1 – 1.6 $mg_{detergent}$/L; SL2 – 3.1 $mg_{detergent}$/L; SL3 – 40 $mg_{solvent}$/L). The use of Principal Component Analysis (PCA) allowed the visualization of the main effects caused by the toxics, by clustering the samples according to its operational phase, exposure or recovery. The morphological parameters showed to be sensitive enough to detect the operational problems even before the COD removal efficiency decreased. Its high loadings in the plane defined by the first and second principal components, which gathers the higher variability in datasets, express the usefulness of monitor the biomass morphology in order to achieve a suitable control of the process. PCA defined a new latent variable t[1], gathering the most relevant variability in dataset, that showed an immediate variation after the toxics were fed to the reactors. t[1] varied 262, 254 and 80%, respectively in SL1, SL2 and SL3. Once more, the high weights of the morphological parameters associated with this new variable express its influence in shock load monitoring and control, and consequently in operational problems recognition.

Keywords
Detergent, Principal Component Analysis, Quantitative Image Analysis, Solvent, Toxic Shock Load

INTRODUCTION

Stable operation of high-rate anaerobic reactors is an essential but difficult task because of the complicated nature of the anaerobic process itself. Monitoring and control are therefore extremely important to improve process robustness by detecting disturbances leading to abnormal process operation. In this context, identification of process variables potentially useful as early alert detectors of instability has major relevance. Industrial wastewaters that are treated by anaerobic digestion processes are frequently affected by temporary toxic exposures. Detergents (Gavala and Ahring, 2002) and solvents (Enright et al., 2005), from cleaning stages, are some of that compounds that can deteriorate the performance of those processes. The earlier a potential fault is detected, the less severe its influence will be and the corresponding corrective action will be more constrained.
An important factor for the efficient operation of anaerobic processes, extensively studied in the last decade, is the recognition of parameters that could be used for monitoring and control of the process. Parameters in the solid phase are not often used for automatic monitoring and control since they usually need manual operations, and are usually qualitative and inaccurate. Therefore, so far, parameters used for control have been limited to indicators of the liquid and the gaseous phases, such as pH, Volatile Fatty Acids (VFA), alkalinity, COD concentrations, carbon dioxide, methane and hydrogen contents in the biogas as well as biogas production (van Lier et al., 2001). In this framework, quantitative image analysis techniques emerge as a promising tool to overcome these difficulties, providing quantitative parameters of the solid phase dynamics. Image analysis has become a very important tool with a large field of applications in study of biomass morphology, due to its ability to remove the subjectiveness of human analysis, the possibility to extract quantitative data and avoid tedious and highly time-consuming tasks to human researchers (Amaral, 2003).
Because the experimental approach of integrating reactor performance, physiological and

morphological data may produce correlated and redundant data, a statistical instrument should be applied in order to extract the essential information for process monitoring and fault detection applications. Often, important information lies not in any individual variable but in how the variables change with respect to one another, *i.e.* how they co-vary (Wise and Gallagher, 1996). Data reduction and interpretation can be approached through the application of multivariate statistical methods, such as Principal Components Analysis (PCA). This method allows identifying patterns in data, and expressing them in order to highlight their similarities and differences. PCA is a projection method for analyze data and reduce it from an n-dimensional space to few latent/hidden variables (Lee et al., 2006), while keeping information on its variability. It has been successfully applied to the monitoring of industrial processes (Li et al., 2000; McGregor and Koutodi, 1995) and wastewater treatment processes (Lee et al, 2004; Lee and Vanrolleghem, 2004; Rosen, 2001). Since patterns in data can be hard to find in data of high dimension, where graphical representation is not available, the possibility of grouping the variability in few variables is an important step to visualize and consequently analyze the information. Chemometrics based techniques are tools that can lead chemists to move more efficiently on the path from measurements to information to knowledge (Frank and Kowalski, 1982).
Previously, three organic load disturbances were applied to lab-scale Expanded Granular Sludge Bed (EGSB) reactors (Costa el al., 2007, 2008a). The corresponding effects were monitored by quantitative image analysis, specific methanogenic activity tests and reactor performance. In the present study, the multivariate statistical tool Principal Components Analysis was applied in order to highlight patterns, groups, trends and outliers in the data. In addition, PCA was employed to identify the variables that mostly reflect the shock load effects, and respective operational changes/problems recognition.

METHODS

Dataset. Three shock loads were applied to Expanded Granular Sludge Bed (EGSB) reactors. In the two first shock loads, SL1 and SL2, a detergent was fed to the reactor with a concentration of 1.6 and 3.1 mg/L, respectively (Costa et al., 2007). In the third shock load (SL3) 40 mg/L of solvent was fed to the EGSB reactor (Costa et al., 2008a). In table 1 are summarised the shock loads conditions.

Table 1. Shock loads conditions.

Shock Load	SL1	SL2	SL3
Ethanol	1.5 g_{COD}/L	1.5 g_{COD}/L	1.5 g_{COD}/L
Toxic	Detergent	Detergent	Solvent
Concentration	1.6 mg/L	3.1 mg/L	40 mg/L
Exposure phase	56 hours	222 hours	222 hours
Recovery phase	14 days	12 days	7 days

Variables summarizing the morphological, physiological and performance data obtained during the experiments were grouped to create the datasets used to perform the PCA (Table 2). Four datasets were created, one for each disturbance, and one integrating the data from all shock loads.

Table 2. Variables included in dataset, summarizing the changes occurred during shock loads.

Variable	Name
Reactor Performance Data:	
OLR	Organic Loading Rate
Cdet	Detergent concentration (*datasets 1 and 2*)
Csol	Solvent Concentration (*dataset 3*)
Tox	Toxic Concentration (detergent or solvent) (*dataset 4*)
Eff	Chemical Oxygen Demand (COD) Removal Efficiency
pH	pH
VSS	Effluent Volatile Suspended Solids

Physiological Data:	
SAA	Specific Acetoclastic Activity
SHMA	Specific Hydrogenotrophic Methanogenic Activity
Morphological Data:	
LfA	Total Filaments Length per Total Aggregates Projected Area
TL/VSS	Total Filaments Length per Volatile Suspended Solids
VSS/TA	VSS per Total Aggregates Projected Area (Apparent Granules Density)
>1	Percentage of Aggregates Projected Area with Equivalent Diameter (D_{eq}) ≥ 1 mm
>0.1	Percentage of Aggregates Projected Area within the range 0.1 ≤ D_{eq} (mm) < 1
<0.1	Percentage of Aggregates Projected Area with D_{eq} < 0.1 mm
vsed	Settling Velocity

Principal Components Analysis (PCA). Principal components analysis aims at finding and interpreting hidden complex, and possibly causally determined, relationships between features in a dataset. Correlating features are converted to the so-called factors which are themselves noncorrelated (Einax et al., 1997). PCA modeling, *i.e.*, the approximation of a matrix by a model, defined by variables and a relatively small number of outer vector products, shows the correlation structure of a data matrix X, approximating it by a matrix product of lower dimension (*TxP'*), called the principal components (PC), plus a matrix of residuals (*E*):

$$X = 1 \times \bar{x}' + T \times P' + E$$

where, the term $1 \times \bar{x}'$ represents the variable averages. The second term, the matrix product $T \times P'$, models the structure and the third term, *E*, contains the deviations between the original values and the projections, *i.e.*, the noise. *T* is a matrix of scores that summarizes the X-variables (scores), and *P* is a matrix of loadings showing the influence of the variables on each score. Geometrically, it corresponds to fitting a line, plane, or hyper plane to the data in the multidimensional space, with the variables as axes. The scaling of the variables specifies the length of the axes of this space.
SIMCA-P (*Umetrics AB*) software package was used to perform the Principal Components Analysis. The first step of the analysis consists in the pre-treatment of data by standardization of the variables, *i.e.*, guarantee that each individual variable has about the same range, avoiding that some variables would be more important than others because of scale effects. During this work each variable was autoscaled by:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

where, x_{ij} is the value of the variable j in the sample i, $\bar{x}_j$ and s_j are the mean and the standard deviation of the variable j, respectively, and, z_{ij} is the autoscaled value of x_{ij}. At the end of this standardization, each variable has mean zero and unit standard deviation.
Subsequently, the software iteratively computes one principal component at a time, comprising a score vector t_a and a loading vector p_a. The score vectors contain information on how the samples relate to each other. Otherwise, the loading vectors define the reduced dimension space and contain information on how the variables relate to each other. Usually, few principal components (2 or 3) can express most of the variability in the dataset when there is a high degree of correlation among data.

The criterion used to determine the model dimensionality (number of significant components) was cross validation (CV). Part of data is kept out of the model development, and then are predicted by the model and compared with the actual values. The prediction error sum of squares (PRESS) is the squared differences between observed and predicted values for the data kept out of the model fitting. This procedure is repeated several times until data element has been kept out once and only once. Therefore, the final PRESS has contributions from all data. For every dimension, SIMCA computes the overall PRESS/SS, where SS is the residual sum of squares of the previous dimension. A component is considered significant if PRESS/SS is statistically smaller than 1.0.

RESULTS AND DISCUSSION

Recognition of shock load effects

Apply a chemometric technique such as Principal Component Analysis (PCA) is advantageous when an effective reduction of the multi dimensional space into few components is accomplished, while keeping the variability of the dataset. In this study, three principal components (PCs) in detergent shock loads (SL1 and SL2) and four PCs in solvent shock load (SL3) gathered more than 80% of the total variability in the datasets (Table 3).

Table 3. Total datasets variability contained in the firsts Principal Components.

PC	SL1	SL2	SL3
1	65.5 %	46.3 %	38.1 %
2	14.3 %	23.9 %	23.8 %
3	9.6 %	14.6 %	12.0 %
4	7.5 %	7.0 %	11.2 %
Cumulative	96.9 %	91.7 %	85.1 %

In the score plots of the first and second PCs, t[1] *vs*. t[2] (Fig. 1a,c,e) is observed that the PCA grouped samples according to its operational phase. A cluster encompassing the observations obtained during exposure phase is visible in each score plot. Besides, is clearly observed that a deviation occurred immediately after the shock loads were applied. The inoculum sample, which emerge as an isolated observation, is located far from the first observation during exposure time (see line in Fig. 1a,c,e).
The influence that each measured variable had in each score, is given by its loadings, *i.e.* weighted variables, and respective loading maps (Fig. 1b,d,e). It allows decide which variables are most important for the differences observed between the samples. The interpretation of the loadings is essentially done by looking at what variables have the higher coefficients (positive or negative) on a certain PC. Coupled visualization of score and loading plots (Fig. 1) allows for the detection of the main effects/problems occurred during the shock loads. For example, the main effects caused by SL1 were detected in the morphological parameters. The introduction of the toxic compound in the feeding caused an increase in LfA and TL/VSS parameters and decrease in VSS/TA (Fig. 1b). These results suggest changes at the granules microstructure level with release of filaments and decrease of apparent density (VSS/TA). However, during reactors operation, the COD removal efficiency remained unaffected (Costa et al., 2007).
Increasing the detergent concentration (SL2) caused an immediate decrease in specific acetoclastic activity (SAA) and VSS/TA (Costa et al., 2007). Analysing the Figure 1c is observed that sample 0 (inoculum) is situated in the top of the graph with the higher score in PC2. Simultaneously, the variables with higher influence in PC2, were SAA and VSS/TA (Fig. 1d, p[2]). PC1 distinguished samples during exposure time (positive scores) from samples during recovery phase/inoculum (negative scores) (Fig. 1c, t[1]). Once more, the morphological parameters LfA and >1, were the most sensitive to recognize the shock load (Fig. 1d, p[1]).
In SL3 the isolation of exposure phase samples is not so effective using only the first PC, since it gathered just 38.1% of the dataset variability (Fig. 1e, t[1]). However, analysing PC1–PC2 plane, a cluster encompassing these samples is visible. The granules size distribution (<0.1, >0.1 and >1) show high loadings in PC2 (Fig. 1f, p[2]). Simultaneously, vsed and VSS present high loadings in PC1 (Fig. 1f, p[1]). Therefore, it is possible to say that these were the variables with higher influence in clustering the samples. Thus, although the reactor performance deteriorates only in the last hours of the exposure phase (Costa et al., 2008a), a change in the macrostructure of granules was observed immediately when the shock load was applied. In fact, the % of aggregates projected area with equivalent diameter (D_{eq}) higher than 1 mm decreased from 81 to 53, indicator of granules fragmentation and consequent washout (increase of the effluent VSS).

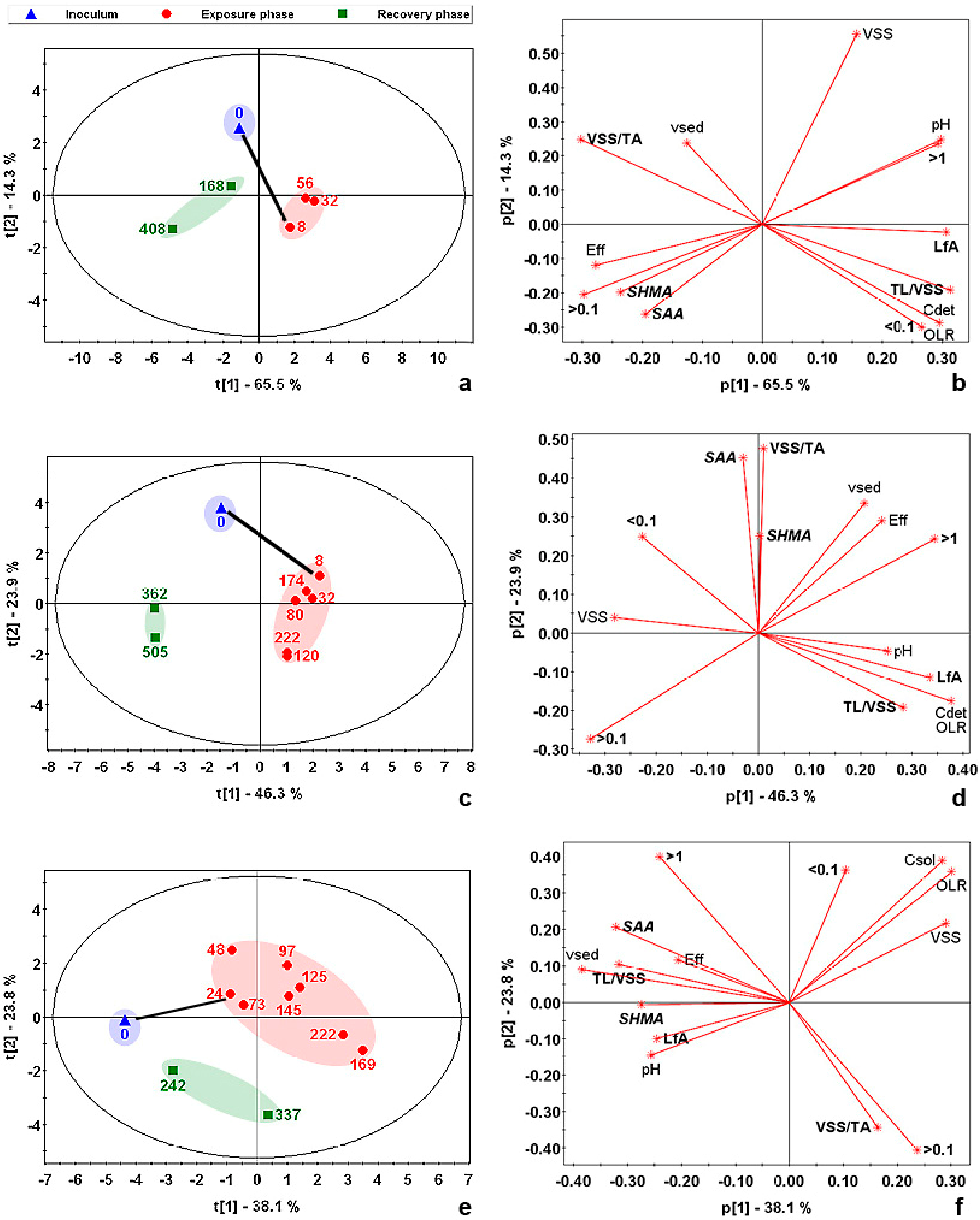

Figure 1. PCA score plot of the first PC (t[1]) versus the second PC (t[2]), in dataset of: (**a**) SL1; (**c**) SL2; and, (**e**) SL3. And, PCA loading plot of the first and second principal components (p[1] vs. p[2]), from dataset of: (**b**) SL1; (**d**) SL2; and, (**f**) SL3.

In the last decade a vast number of methods to monitoring and/or control of wastewater anaerobic digestion processes have been proposed with different parameters as indicators of operational problems (Garcia et al., 2007, Lardon et al., 2005). However, the integration of morphological parameters has not yet been studied, mainly because expeditiously and quantitative information is difficult to obtain. The use of image analysis techniques, previously described by Amaral (2003) and Costa et al. (2007) provides quantitative information about the dynamic evolution of the granules morphology at macro and microstructures levels. The use of PCA illustrates the usefulness of monitoring the granules morphology to detect possible toxic contamination and future operational problems. The early detection of these problems is essential to attain timely control of the process before it evaluates to an irreversible problem. In this work was visible that morphological changes occurred before reactors performance deterioration, proving the sensitivity of the proposed parameters to detect the toxic contaminations.

PCA provides information on the most meaningful parameters, which describes a whole dataset affording data reduction with minimum loss of original information (Helena et al., 2000). During PCA, it was defined a new latent variable, t[1], that includes a weighted sum of performance, physiological and morphological information. This new variable can be used as a warning indicator of operational problems during toxic shock load disturbances. The variable t[1] was calculated for the inoculum and the first sample of exposure phase and the corresponding % of variation was 262, 254 and 80%, respectively in SL1, SL2 and SL3. This result evidenced the high sensitivity of the latent variable to recognize deviations of the normal process operation.

Analyzing the loadings/weights associated with the new latent variable t[1], it is possible to distinguish the variables that most influence the early detection of reactors contaminations. The morphological parameters emerge due to its high loadings in all datasets (Table 4). These results confirm that quantitative morphological parameters should be considered in monitoring and control of high rate anaerobic reactors, especially those based on granular sludge. Similar conclusions were obtained when high-rate anaerobic reactors were subjected to organic loading disturbances (Costa et al. 2008b).

Table 4. Loadings/weights of the variables in datasets associated to the PC1.

Variable	SL1	SL2	SL3	Notes
OLR	0.295	0.377	0.300	Controlled Variable
Cdet	0.295	0.377	-	Controlled Variable
Csol	-	-	0.283	Controlled Variable
Eff	-0.278	0.241	-0.206	
pH	0.297	0.253	-0.256	
VSS	0.158	-0.280	0.290	
<0.1	0.265	-0.227	0.104	Morphological Variable
>0.1	-0.297	**-0.329**	0.236	Morphological Variable
>1	0.293	**0.345**	-0.241	Morphological Variable
SAA	-0.194	-0.030	**-0.321**	
SHMA	-0.236	0.003	-0.275	
LfA	**0.306**	**0.336**	-0.246	Morphological Variable
VSS/TA	**-0.302**	0.009	0.164	Morphological Variable
TL/VSS	**0.313**	0.283	**-0.316**	Morphological Variable
vsed	-0.126	0.207	**-0.384**	

Differentiate the shock loads

A PCA in a dataset integrating all available information was performed in order to highlight differences between the shock loads. Watching at Figure 2a three clusters, one for each shock load, can be perfectly distinguished. The cluster encompassing the SL2 samples is isolated from the others. Effectively, SL2 caused the most negative effects to the anaerobic granular sludge, since it was the only one where the COD removal efficiency decreased significantly during the exposure phase.

The score and loading plots of PC1 and PC2, t[1] *vs.* t[2] and p[1] *vs*. p[2] (Fig. 2), show the variables with higher influence in each shock load. SL1 was characterized by an increase in TL/VSS and LfA. The decrease of Efficiency (Eff) and SAA and increase of granules density describe SL2. Regarding to SL3 it was categorized mostly by the granules D_{eq} ranges >1 and >0.1, sign of granules fragmentation.

Searching for possible correlations between variables, it is possible to observe a high positive correlation between the total filaments (TL/VSS) and the dynamic of filaments per area of aggregates (LfA) (Fig. 2b). This was already postulated by Costa et al (2008b), suggesting that the granules microstructure stabilization, by locking the filaments inside the aggregates, play a more important role in the maintenance of a high efficiency than granules macrostructure/size stabilization.

During the shock loads was observed that LfA increased 3, 5, and 2 days before effluent volatile suspended solids, respectively in SL1, SL2, and SL3. It was hypothesized that LfA could be an early-warning indicator of washout events (Amaral et al., 2004, Costa et al., 2007, 2008a). In Figure 2b is visible that LfA and VSS are inversely proportional, enhancing the hypothesis that LfA increases before VSS, decreasing afterwards when VSS increases.

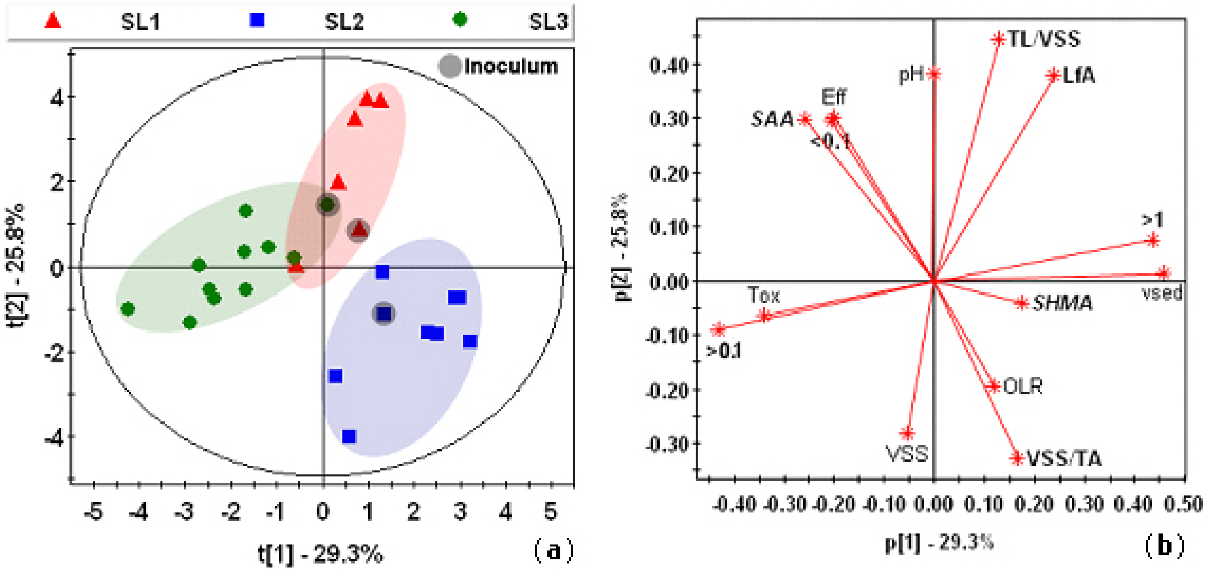

Figure 2. PCA in dataset integrating data from all shock loads: (**a**) score plot of the first PC (t[1]) versus the second PC (t[2]); and, (**b**) loading plot of the first and second principal components (p[1] vs. p[2]).

CONCLUSIONS

Principal Component Analysis was performed in three datasets gathering morphological, physiological and Expanded Granular Sludge Bed reactor performance information obtained during three toxic shock loads. It was demonstrated that the use of a multivariate statistical tool was appropriate to visualize and isolate the main effects caused by the detergent and solvent shock loads.

The proposed morphological parameters proved to be more sensitive to detect the toxic contaminations than the normal operating parameters, such as Chemical Oxygen Demand removal efficiency. In shock load 1 (1.6 $mg_{detergent}.L^{-1}$) and 3 (40 $mg_{solvent}.L^{-1}$), although the reactors performance seemed to be unaffected by the toxics concentration and exposure time, changes in micro and macrostructure of the granules were observed. In shock load 2 (300 $mg_{detergent}.L^{-1}$) the morphological changes were detected in the morphological parameters before reactor efficiency decreased. The new latent variable t[1], defined as an weighted sum of all variables included in the dataset, showed a variation of 262, 254 and 80 %, respectively in SL1, SL2 and SL3. The high loadings/weights of the morphological parameters enhanced the need to monitor the anaerobic digestion process solid phase in order to achieve an effective and feed forward control.

Integrating all information in a single dataset allowed the differentiation of the several shock loads. The shock load 2 (SL2) had the most negative effects to the anaerobic granular sludge. Consequently, the observations in SL2 were grouped in a cluster opposite to the clusters relating SL1 and SL3.

REFERENCES

Amaral A.L. (2003). *Image Analysis in Biotechnological Processes: Application to Wastewater Treatment.* PhD thesis, University of Minho, Portugal (http://hdl.handle.net/1822/4506).

Amaral A.L., Pereira M.A., da Motta M., Pons M.-N., Mota M., Ferreira E.C., and Alves M.M. (2004). Development of image analysis techniques as a tool to detect and quantify morphological changes in anaerobic sludge: II. Application to a granule deterioration process triggered by contact with oleic acid. *Biotechnology and Bioengineering,* **87**(2): 194-199.

Costa J.C., Abreu A.A., Ferreira E.C., and Alves M.M. (2007). Quantitative image analysis as a diagnostic tool for monitoring structural changes of anaerobic granular sludge during detergent shock loads. *Biotechnology and Bioengineering,* **98**(1): 60-68.

Costa J.C., Moita I., Ferreira E.C., and Alves M.M. (2008a). Morphology and physiology of anaerobic granular sludge exposed to organic solvents. *Bioresource Technology,* (submitted).

Costa J.C., Moita, I., Abreu, A.A., Ferreira, E.C., and Alves, M.M. (2008b). Advanced Monitoring of High Rate Anaerobic Reactors through Quantitative Image analysis of granular sludge and Multivariate Statistical Analysis. *Environmental Science & Technology*, (submitted).

Einax J.W., Zwanziger H.W, and Geiss S (1997). *Chemometrics in Environmental Analysis*. Weinheim: VCH.

Enright A.-M., McHugh S., Collins G. and O'Flaherty V. (2005). Low-temperature anaerobic biological treatment of solvent-containing pharmaceutical wastewater. Water Research, 39(19): 4587-4596.

Frank I.E., and Kowalski B.R. (1982). Chemometrics. *Anal. Chem.*, 54, 232R-243R.

Garcia C., Molina F., Roca E.; and Lema J.M. (2007). Fuzzy-based control of an anerobic reactor treating wastewaters containing ethanol and carbohydrates. *Ind. Eng. Chem. Res.*, 46 (21): 6707-6715.

Gavala H.N., and Ahring B.K. (2002). Inhibition of the anaerobic digestion process by linear alkylbenzene sulfonates. Biodegradation 13:201–209.

Lardon L., Puñal A., Martinez J.A., and Steyer J.P. (2005). Modular expert system for the diagnosis of operating conditions of industrial anaerobic digestion plants. *Water Science & Technology*, 52 (1-2): 427-433.

Lee D.S., and Vanrolleghem, P.A. Adaptive consensus principal component analysis for on-line batch process monitoring. *Environmental Monitoring and Assessment*, 92: 119-135.

Lee D.S., Lee M.W., Woo S.H., Kim Y.-J., and Park J.M. (2006). Multivariate online monitoring of a full-scale biological anaerobic filter process using kernel-based algorithms. *Ind. Eng. Chem. Res.*, 45: 4335-4344.

Lee J.M., Yoo C.K., Choi S.W., Vanrolleghem P.A., and Lee I.B. (2004). Nonlinear process monitoring using kernel principal component analysis. *Chemical Engineering Science,* 59: 223-234.

Li W., Yue H., Valle-Cervantes S., and Qin S.J. (2000). Recursive PCA for adaptive process monitoring. *Journal of Process Control,* 10: 471-486.

MacGregor JF, and Koutodi M. (1995). Statistical process control of multivariate process. *Control Engineering Practice,* 3(3): 403-414.

Rosen C. A. (2001). *Chemometric approach to process monitoring and control with applications to wastewater treatment operation.* Ph.D thesis, Lund University, Sweden.

Van Lier J.B., Tilche A., Ahring B.K., Macarie H., Moletta R., Dohanyos M., Hulshoff Pol L.W., Lens P., and Verstraete W. (2001). New perspectives in anaerobic digestion. *Wat. Sci. Technol.,* 43(1): 1-18.

Wise B.M., and Gallagher N.B. (1996). The process Chemometrics approach to process monitoring and fault detection. *J. Proc. Cont.,* 6(6): 329-348.