Presented at Poster Session PT03 – Process Analysis
11th IWA World Congress on Anaerobic Digestion, 23-27 September 2007, Brisbane, Australia

Integration of morphological and physiological data through Principal Component Analysis to identify the effect of organic overloads on anaerobic granular sludge

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

IBB – Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, University of Minho, 4710-057 Braga, Portugal.
E-mail: *carloscosta@deb.uminho.pt; madalena.alves@deb.uminho.pt; ecferreira@deb.uminho.pt*

Abstract Morphological parameters, obtained by quantitative image analysis techniques, together with physiological and reactor performance data were inserted in principal components analysis (PCA) to detect operational problems and control of high rate anaerobic reactors during organic overloads. Four lab-scale Expanded Granular Sludge Blanket reactors were used to performed organic overloads of 18 $kg.m^{-3}.day^{-1}$ (R1 – HRT of 8h; and, R2 – HRT of 2.5h) and 50 $kg.m^{-3}.day^{-1}$ (R3 - fed for 3 days; and, R4 - fed for 16 days). The application of PCA allowed the visualization of the main effects caused by the organic overloads. The first Principal Component (PC) extracted, in each shock load, retains enough information to group observations in agreement with operational conditions (normal or overload). The variables from quantitative image analysis presented high loadings, suggesting that might play an important role in organic overloads control.

Keywords Anaerobic granular sludge; methanogenic activity; principal component analysis; quantitative image analysis; organic overload.

INTRODUCTION

Biological wastewater treatment plants are normally designed with reference to a nominal operating condition, in which the loading rate is assumed to be constant in time. However, in practice this steady-state assumption is seldom met and in fact the process is subject to wide fluctuations, both in flow and organic loading rate, which often result in performance degradation or even plant failure (Muller et al., 1997).

Integration of reactor performance, physiological and morphological data, in monitoring of anaerobic processes produces a set of correlated and redundant data. These data must be compressed to retain only the essential information. 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).

The problem of data reduction and interpretation can be approached through the application of multivariate statistical methods, such as, Principal Components Analysis (PCA). Multivariate statistical techniques have been successfully applied for monitoring and modeling chemical/biological processes (Lee et al., 2006). PCA is a projected method for analyzing a historical reference distribution of the measurement trajectories from past successful operations, in a reduced latent vector space and comparing the behaviours of new operations to this reference distribution.

This work intends to integrate new information, supplied by quantitative image analysis techniques, about changes occurred in anaerobic granular sludge morphology during organic overloads, with physiological and reactor performance data. Afterwards, the application of PCA to databases reviewing the main parameters that summarize the operational conditions and changes occurred, was performed to identify operational problems and find correlations between samples and/or variables.

METHODS

Principal components (PC) analysis are aimed at finding and interpreting hidden complex, and possibly causally determined, relationships between features in a data set. Correlating features are converted to the so-called factors which are themselves noncorrelated (Einax et al, 1997). PCA modelling shows the correlation structure of data matrix X, approximating it by a matrix product of lower dimension (TP'), called the principal components, plus a matrix of residuals (E).

$$X = 1 * \overline{X} + TP' + E \quad (1)$$

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. **T** is a matrix of scores that summarizes the X-variables, and **P** is a matrix of loadings showing the influence of the variables on each score. **E** is a matrix of residuals; the deviations between the original values and the projections. The residual standard deviation (RSD) can be computed for observations and variables. The RSD of an observation (rows in E) is also called the observation distance to the PC model (DModX). The RSD of a variable relates to the variable relevance in the PC model. SIMCA-P (Umetrics AB) software package was used to perform the PCA; it iteratively computes one principal component at a time, comprising a score vector ta and a loading vector pa. 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 each other. Usually, a few PC (2 or 3) can express most of the variability in the database when there is a high degree of correlation among data. The criterion used to determine the model dimensionality (number of significant components) is 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

No significant effects were detected in performance of reactor R1. In reactor R2 the efficiency dropped from 94 to 72 %, although, no significant changes occurred in the Specific Methanogenic Activity (SMA). Granules erosion was observed in R1 (area of granules with Deq>1mm decreased from 90 to 70%). While in R2 the washout phenomenon was immediate in R1 this was more severe (Costa et al., 2007a). Reactors R3 and R4 showed decreases of efficiency from 90 to 30%. The SMA was inhibited during adapting phase (24h). Afterwards, in R3, the SMA was recovered. Relatively to R4, although SMA in H2/CO2 recovered, the SMA in acetate was reduced during the exposure time, meaning that acetoclastic bacteria were inhibited. Extend of exposure time in R4 provoked acidogenesis inhibition after 144h. Granules fragmentation, translated by 45% reduction in area of aggregates with Deq>1mm, was observed. Filaments release and washout of biomass occurred. In R4 a new dynamic in granules structure and biomass washout was obtained; only the denser granules remained inside the reactor. Small and fluffier granules were present at the end of shock loads (Costa et al., 2007b).

The database created, reviews the most influent variables in: operation conditions (organic loading rate – OLR, and, hydraulic retention time – HRT); reactor performance (COD removal efficiency – ReEf, pH, and effluent VSS); specific acetoclastic activity – SAA, and, specific hydrogenotrophic methanogenic activity – SHMA); and, quantitative image analysis (total area of aggregates distribution by D_{eq} ranges (0.01-0.1 mm (%>0.01), 0.1-1 mm (%>0.1) and >1 mm (%>1)), LfA, VSS/TA and TL/VSS). All variables were autoscaled to unit variance.

The score (**ti**) of an observation (**i**) on a principal component (**PC**) j is the weighted sum of the original variables (xi). The weights (pi) are called the loadings of the variables on that **PCj**. The loading of a variable is related to its variation (Massart and Heyden, 2005).

$$tiPCj = \sum piPCjxi \quad (2)$$

R1 (OLR = 18 kg m^{-3} day^{-1}; HRT = 8 h)

PC1-PC2 plane split the recovery phase (96, 192 and 235) from the inoculum (0), and, shock load (8, 24 and 72) samples (Fig. 1a). The extraction of 3 PC's explained 84.2 % of the total variability of data. 44.4 % of the total variability is present in PC1. Concomitantly, the high variations of LfA, TL/VSS, VSS/TA and SHMA occurred when reactor is back to the initial operation conditions, explaining its high loadings in PC1 (Fig. 1b, p[1]). Second and third PC explained more 20.5 and 19.3 % of the total variability, respectively. The variables that most influenced the scores in PC2 (Fig. 1a, t[2]), i.e. with higher loadings, are VSS, pH, %>0.1 and %>1 (Fig. 1b, p[2]). Therefore, the increase in effluent VSS, coupled with changes in aggregates projected area distribution by Deq, during shock load, specify the direction of second highest variance in organic shock load.

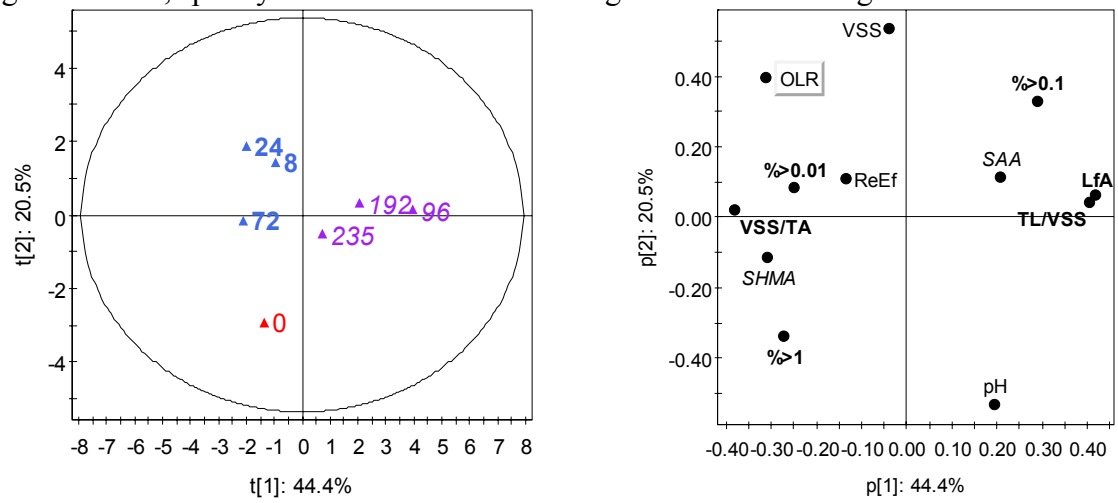

Figure 1 – (**a**) Score map t[1] vs. t[2], and, (**b**) Loading map p[1] vs. p[2], of organic shock load.

R2 (OLR = 18 kg m^{-3} day^{-1}; HRT = 2.5 h)

PCA was performed to data compression by means of a projection of the original variables onto a lower dimensional space. 86 % of the total variability present in database was contained in 3 PC's. PC1–PC2 plane (Fig. 2a) encloses the most information of all planes that can be drawn through the data in the multi-dimensional space. Analyzing this plane (67.1% of the total variability), three "clusters" can be distinguished. First "cluster" is located in the upper part of the plot, refer to sample 0 (inoculum) with high score in PC2 (22.5 % of variance) (Fig.2a, t[2]. The variables that most influence this PC are %>1, %>0.1, SHMA, ReEf and VSS (Fig. 2b, p[2]). Second and third cluster refers to observations with negative (samples 8, 24 and 72) and positive (samples 96, 160 and 232) scores in PC1 (Fig. 2a, t[1]), respectively from shock load and recovery phases. The variables with higher loadings in PC1 (Fig. 2b, p[1]), i.e. responsible for grouping shock load and recovery observations are HRT, pH, SAA, OLR, LfA and TL/VSS. Therefore, it can be concluded that the highest changes (variation) caused by the decrease of HRT and consequent increase in OLR were mainly detected in SAA, LfA, TL/VSS and pH parameters, followed by ReEf and VSS variables. Focusing the analysis on how variables co-vary between themselves, it is observed (Fig. 2b) that: the HRT decrease will cause the pH decrease (variables with similar coordinates); and, the increases of LfA, TL/VSS and VSS. This is visualized by searching for variables with similar and symmetric coordinates.

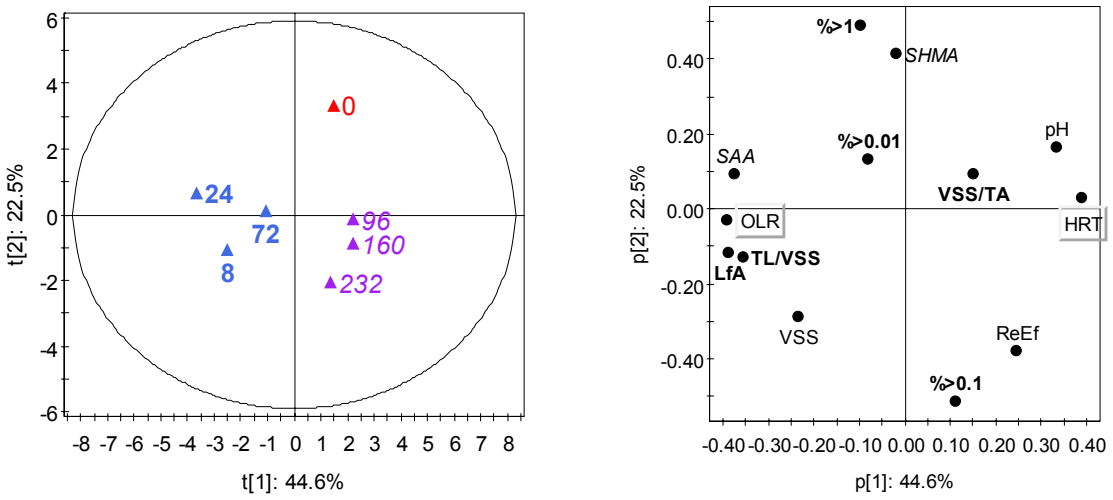

Figure 2 – (a) Score map t[1] vs. t[2], and, **(b)** Loading map p[1] vs. p[2] of hydraulic shock load.

R3 (OLR = 50 kg m^{-3} day^{-1}; exposure time = 3 days)

The first PC (Fig. 3a, t[1]), translating 56.6% of the total variability in database allow the partition of observations in agreement with operational conditions. The samples 0 (inoculum), 96, 192, and 235 (recovery phase), correspond to normal conditions, and show positive scores. The samples 8, 24, and 72 correspond to overload conditions, showing negative scores. Considering the loadings/coefficients, it was observed that almost all variables contributes to the variability in PC1, with relevance to pH, ReEf, VSS, LfA and TL/VSS, with coefficients higher than 0.30 (Fig. 3b, p[1]). The coupled analysis of scores and loadings on PC1, suggest that the quantitative image analysis techniques provide powerful parameters to detection and differentiation between normal and shock load conditions. The second PC (Fig. 3a, t[2]) explained 21.3 % of the total variability. The differences between inoculum and recovery phase samples are translated in this component. It tells that if almost all the parameters return to their initial values after the overload, it does not necessarily indicate that the effects of the overload were totally solved. Another major variance was detected between shock load observations. It is observed large deviation in sample 24. It corresponds to changes in specific methanogenic activity and image analysis results. The variables with high loadings on PC2 (Fig. 3b, p[2]) are essentially the image analysis parameters of the distribution by % of aggregates in each D_{eq} range, 0.01–0.1 mm, 0.1–1 mm, and >1mm (%>0.01, %>0.1 and %>1, respectively), and SAA and SHMA. Therefore, as expected and observed by the scores on PC2, it can be concluded that the major effects discriminating between inoculum and recovery phase samples, and involving shock load samples were caused/detected in the image analysis and methanogenic activity parameters. Focusing the analysis on the visualization of loading map [p1] vs. [p2] (Fig 3b), it allows the detection of the parameters most affected by the shock load. The variable that represents the change in operational conditions (OLR), although inversely proportional, is highly correlated, with ReEf, pH, and %>1, meaning that the increase of OLR will cause the decrease of COD removal efficiency and pH, and increase granules fragmentation/erosion (decreasing the % of granules with D_{eq} >1mm). The variables OLR, VSS, and %>0.1 are directly proportional, denoting that the increase of OLR will cause an increase in effluent VSS and % of granules with D_{eq} in the range 0.01–0.1 mm. Applying the same reasoning, other relations can be obtained from the loading map (Fig 3b). It is observed that LfA is highly correlated with TL/VSS. Simultaneously it is inversely proportional to SAA and SHMA and to VSS/TA, suggesting that the increase of LfA will cause the decrease of methanogenic activities and granules density.

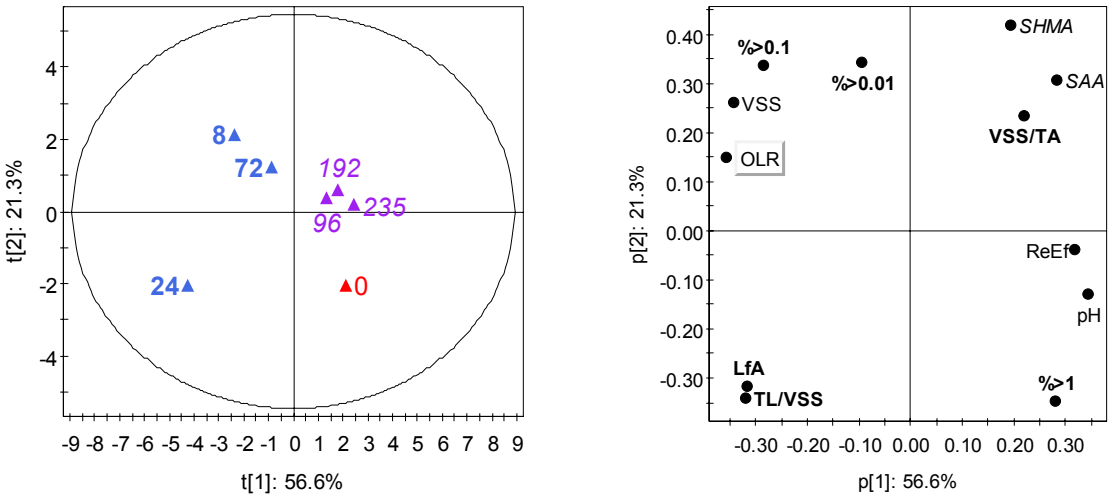

Figure 3 – (**a**) Score map t[1] vs. t[2], and, (**b**) Loading map p[1] vs. p[2].

R4 (OLR = 50 kg m^{-3} day^{-1}; exposure time = 16 days)

Increasing the exposure time it is expected that the database variability increases. Also, the negative effects provoked by increasing OLR were emphasized, causing slight alterations in loadings and respective loading maps of PCA. The first PC extracted explains only 50.6% of the total variability in the database (Fig. 4a t[1]). The variance between normal and overload observations is translated in this component. The variables with high loadings on PC1 (Fig. 4b, p[1]), therefore the most relevant for explain the variability between normal and overload samples, are COD removal efficiency (ReEf), SAA, LfA, effluent VSS, and distribution of aggregates projected area by D_{eq} ranges (%>1 and %>0.1). The increase of exposure time, lead to acidogenesis inhibition causing the pH raise, since VFA accumulation decreased. This fact might explain the lost of importance of pH to detect changes caused by this shock load in relation to R3, where there was not enough time to acidogenesis inhibition. Image analysis parameters play an important role in the detection of shock loads, especially in the first, and more important, hours of exposure.

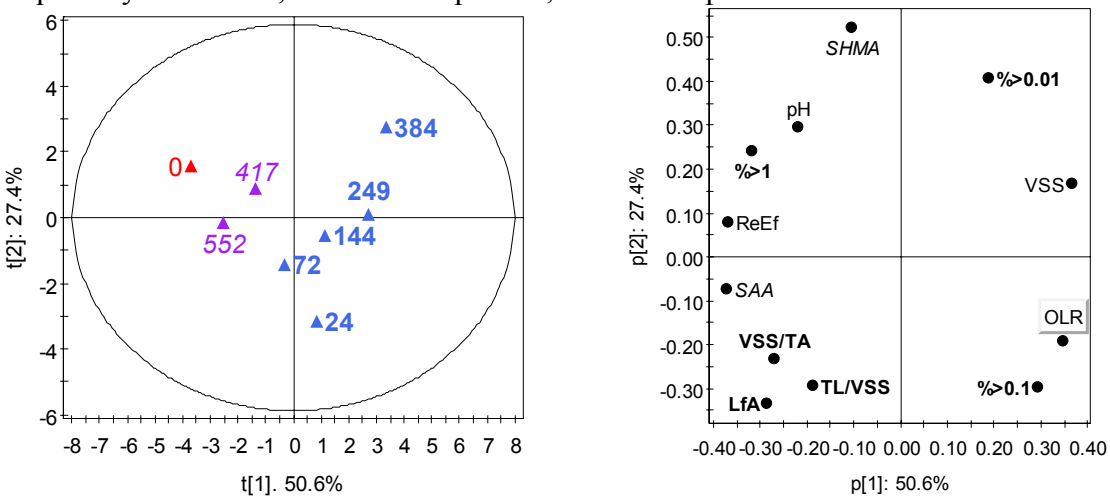

Figure 4 – (**a**) Score map t[1] vs. t[2], and, (**b**) Loading map p[1] vs. p[2].

Analyzing the plot, involving the scores of the two PC's, t [1] vs. t [2] (Fig. 4a), visualizing the plane containing the highest variance in the database (78%), is possible to identify two main clusters. The effects of increase in OLR may be divided in two different periods. The first one (until t = 144 h), as expected, showed similar trends as R3. The main effects caused by extend exposure time was acidogenesis inhibition and a new dynamic involving morphological parameters

controlling washout, that characterized the second period of overload. The main cluster recognized in this plane refer to: first period of shock load (samples 24, 72 and 144) characterized by the decrease of COD removal efficiency and granules erosion/fragmentation; second period of shock load (samples 249 and 384) characterized by acidogenesis inhibition and high washout of biomass (samples 384 and 249). The other cluster include inoculum and recovery phase (samples 0, 417 and 552) typifying the normal operational conditions. The variables that most influence the scores on PC2 are SHMA, %>0.01, %>0.1, TL/VSS, pH, and LfA (Fig. 4b, p[2]). Analyzing the results, together with the score map [t1] vs. [t2] (Fig. 4a), it might be said that this variables were the most suitable to monitor the adjustments occurred in the microstructure of granules, simultaneously with acidogenesis inhibition (samples 249 and 384). These facts corroborate that this inhibition might be connected with the washout, caused by granules erosion, of consortia responsible for this step.

CONCLUSIONS

The visualisation of morphological structural changes of anaerobic granular sludge by quantitative image analysis and principal component analysis techniques might bring relevant information to control EGSB reactors under severe organic shock loads.

The extraction of two principal components was enough to group observations as normal or overloaded. The loading maps [p1] vs. [p2], gathers the most relevant information to detect changes/effects caused by severe organic loading rates. It is relevant to note that the quantitative image analysis parameters, specially LfA, TL/VSS, and granules area distribution by Deq ranges, might have an important role in earlier recognition of operational problems. Together with COD removal efficiency, pH and VSS, they presented the highest loadings, i.e. large variance, to separate the scores of first hours of operational instability.

ACKNOWLEDGEMENTS

We grateful acknowledge the financial support to J.C. Costa through the grant SFRH/BD/13317/2003 and the project POCI/AMB/60141/2004, from the Fundação para a Ciência e a Tecnologia (Portugal).

REFERENCES

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

Costa J.C., Moita I., Abreu A.A., Ferreira E.C. and Alves M.M. (2007a). Image analysis and multivariate statistical techniques as instruments to identify changes in anaerobic granular sludge during organic overloads: Part I – Influence of hydraulic retention time. (submitted).

Costa J.C., Moita I., Abreu A.A., Ferreira E.C. and Alves M.M. (2007b). Image analysis and multivariate statistical techniques as instruments to identify changes in anaerobic granular sludge during organic overloads: Part II – Influence of exposure time. (submitted).

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.

Massart, D.L. and Vander Heyden, Y. (2005). From tables from visuals: principal component analysis, Part 2. Practical Data Handling. *LC•GC.* **18**(2), 84-89.

Muller, A., Marsili-Libelli S., Aivasidis A., Lloyd T., Kroner S. and Wandrey C. (1997). Fuzzy control of disturbances in a Wastewater treatment process. *Wat. Res.* **31**(12), 3157-3167.

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.