KV_C003
Villez, K., Lee, D. S., Rosén, C., Vanrolleghem, P. A. (2006). Comparison of linear and non-linear PLS methods for soft-sensing of an SBR for nutrient removal. Proceedings of the 3rd Biennial meeting of the International Environmental Modelling and Software Society (iEMSs2006), Burlington, VT, USA, July 9-12, 2006, appeared on CD-ROM.

Comparison of linear and non-linear PLS methods for soft-sensing of an SBR for nutrient removal

K. Villez[a], **D.S. Lee**[b], **C. Rosen**[c] **and P.A. Vanrolleghem**[a,d]

[a] *BIOMATH, Department of Applied Mathematics, Biometrics and Process Control, Ghent Univeristy, Coupure Links 653, B-9000 Gent, Belgium*

E-mail: Kris.Villez@biomath.UGent.be

[b] *Department of Environmental Engineering, Kyungpook National University, Sankyuk-dong, Buk-gu, Daegu 702-701, Korea*

[c] *IEA: Department of Industrial Electrical Engineering and Automation, Lund University, LTH, Box 118, SE-221 00, Lund, Sweden*

[d] *modelEAU: Département de Génie Civil, Pavillon Pouliot, Université Laval, Québec, G1K 7P4, QC, Canada*

Abstract: Despite of promising results in research, advanced control strategies fail to gain trust in wastewater treatment practice. Due to the sensitivity of the biological processes to disturbances, operators are often unable to find the causes of faults due to the lack of effective real-time on-line monitoring. Strategies for on-line monitoring are therefore essential to enhance biological process control. Therefore, a suitable multivariate soft-sensor is desired for fault detection and control for a pilot-scale sequencing batch reactor (SBR) system to allow effluent quality to be estimated well before off-line analysis is finished. For this purpose, several multivariate methods are available, including (linear) partial least squares (PLS), Neural Net PLS (NNPLS) and Kernel PLS (KPLS). While non-linear extensions of PLS such as NNPLS require fitting of non-linear functions, KPLS does not. KPLS is based on a non-linear transformation of the process data, followed by the fitting of a linear PLS model between the transformed inputs and outputs. PLS, NNPLS and KPLS were compared regarding their ability to predict effluent quality data and their computational requirements. While (linear) PLS and NNPLS lead to acceptable prediction, KPLS results in poor model performance. Moreover, the computational requirement of KPLS were large compared to PLS and NNPLS. When comparing PLS and NNPLS to each other, it was found that NNPLS leads to the best possible prediction given the experimental data set, while the extra computational requirements are minimal.

Keywords: Partial Least Squares (PLS); Neural Net PLS (NNPLS); Kernel PLS (KPLS); On-line process monitoring and control; Biological wastewater treatment plants; Supervisory control

1. INTRODUCTION

SBR technology has received increasing attention in the framework of wastewater treatment in the past decades. One of the most attractive features of such systems is their high degree of operational flexibility. Inspired by the increasing amounts of data that can be collected, PCA- and PLS-based tools have been introduced for data dimension reduction and process monitoring since the works of Nomikos and MacGregor [1994] and Wold *et al.* [1998]. Applications of PLS to continuous activated sludge systems can be found in Teppola *et al.* [1997] and Mujunen *et al.* [1998] and Lee *et al.* [2005].

A PLS-based approach to effluent quality prediction of batch processes for wastewater treatment is presented in this work. Three different PLS-based models are evaluated for prediction of effluent quality of a pilot-scale SBR for nutrient removal. These techniques include (linear) partial least squares (PLS), neural net partial least squares (NNPLS) and kernel partial least squares (KPLS). In section 2, a short description of the used data set is given, next to an overview of the applied methods. In section 3, results are shown, followed

by section 4 providing the discussion. Section 5 holds conclusions and suggestions for further research.

2. MATERIALS AND METHODS

2.1 Data

The data was derived from a pilot-scale SBR for nutrient removal from December 16th, 2003 until May 12th, 2005. A technical description of the setup and the synthetic influent can be found in Insel *et al.* [2004]. The details of the time-based control scheme that was applied are described in Sin *et al.* [2005]. The complete dataset consists of 1587 observations (batches).

The data of the on-line sensors were used as predictors (inputs). For each batch, this corresponds to the (6) trajectories of the volume, temperature, dissolved oxygen (DO), pH, oxido-reduction potential (ORP) and conductivity. Each trajectory consists of 300 measurements, taken with 1-minute intervals in the first 5 hours of each batch. The last hour of each batch was not taken into account as changed sensor positions prevent straightforward interpretation of the (non-mixed) settling phase data. The outputs or responses consist of the effluent concentrations of total nitrogen (TN), nitrate nitrogen (NO_3^-) and total phosphorous (TP).

The data set was split into a model (calibration) and test (validation) set, representing respectively 80% and 20% of the dataset. As the process was subjected to significant changes in operation during the studied timeframe, the observations were randomly assigned to one of the sets.

2.2 Data unfolding and scaling

The process data of a batch process is of 3-dimensional nature where the 3 axes represent batch number, sensor or variable number and the batch runtime. For reasons of interpretation Gurden *et al.* (2001) prefer the use of N-PLS models over Unfolding PLS (U-PLS). However, this preference is constrained to the existence of a multi-linear structure in the data, which is not evident in our case. Therefore, Unfolded PLS (U-PLS) was selected and performed as described in Nomikos and MacGregor (1995).

2.3 Partial Least Squares

PLS is a tool aimed at a dimension reduction of the inputs, denoted **X**, by extraction of latent variables which are maximally correlated with the outputs, **Y**, while maximizing the amount of variance captured in the input matrix (**X**).

In summary, a PLS model is defined by the following set of equations:

$$\mathbf{X} = \sum_{i=1}^{c} \mathbf{t}_i \mathbf{p}_i^T + \mathbf{E} \quad (1) \qquad \mathbf{Y} = \sum_{i=1}^{c} \mathbf{u}_i \mathbf{q}_i^T + \mathbf{F} \quad (2)$$

where **X** and **Y** are the scaled input and output matrices and c defines the number of latent variables, being the so-called meta-parameter of the PLS model. $\mathbf{p}_i$ and $\mathbf{q}_i$ represent the loadings of the corresponding latent variables in the input and output space respectively, while **E** and **F** represent the residuals in the input and output space respectively. Linear regression of vector $\mathbf{u}_i$ on $\mathbf{t}_i$ results in the following inner relation:

$$\mathbf{u}_i = b_i \mathbf{t}_i + \mathbf{h}_i \quad (3)$$

where b_i is the regression coefficient obtained by minimisation of the residuals $\mathbf{h}_i$. In this work, the NIPALS (Nonlinear Iterative Partial Least Squares) algorithm as presented in Geladi and Kowalski [1986] was used.

2.4 Neural Net Partial Least Squares

Where PLS is limited by its ability to extract linear relations only, NNPLS is able to extract non-linear relationships by fitting a 3-layer (1 hidden layer) neural network between the respective input and output scores. While equations (1) and (2) remain the same, the inner relation is now defined as:

$$\mathbf{u}_i = \mathrm{FBNN}(\mathbf{t}_i) + \mathbf{h}_i \quad (4)$$

where FBNN(.) represents the fitted feed-forward back-propagation neural network (FBNN) and $\mathbf{h}_i$ the residuals. As such, NNPLS is a tool for non-linear modelling when faced with collinear inputs [Wold, 1989].

2.5 Kernel Partial Least Squares

KPLS is another PLS-based method that is suited to model non-linear systems. KPLS is based on the non-linear kernel transformation of the input data, followed by linear PLS modelling. The nonlinear mapping consists of computing the kernel matrix:

$$\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{d}\right) \quad (5)$$

where d represents the kernel width, a meta-parameter or tuning parameter of the resulting regression model, and the vectors $\mathbf{x}_i$, $\mathbf{x}_j$, represent input observations where i and j indicate the sample number. The second step in the procedure consists of regression of the output variable onto the resulted kernel matrix. The PLS model is then derived by means of PLS regression of the outputs onto the transformed inputs. More details and

justification of the Kernel Partial Least Squares can be found in Schölkopf *et al.* [1999] and Rosipal and Trejo [2001].

3. RESULTS

3.1 Partial Least Squares

Figure 1 shows the results of PLS and NNPLS regression of total nitrogen (TN), nitrate (NO_3^-) and total phosphorous (TP) onto the process data. The figure shows the sum of squared prediction errors (RRMSE) over the validation data set for the first 15 latent variables (LV's). Based on the first graph (TN), a 7-LV model is selected for TN prediction. Figure 2 shows the original and estimated data for the validation dataset. As can be seen, the model is able to capture the overall long-term trend in the dataset but fails to provide a fully reliable estimate of TN values.

Similar models were made for nitrate (NO3) and phosphorous (TP) prediction. 7 LV's were retained based on the second graph (NO3) in Figure 1. Figure 2 shows the original and predicted data in the validation dataset. The model could capture trends well in the model data set (not shown) but is seen to fail quite often in the validation dataset.

On the basis of the third graph in Figure 1 an 8-LV model was selected for TP. Original and predicted data in the validation data set are shown in Figure 3. The model is captures major trends in the data but the obtained prediction may not be satisfying.

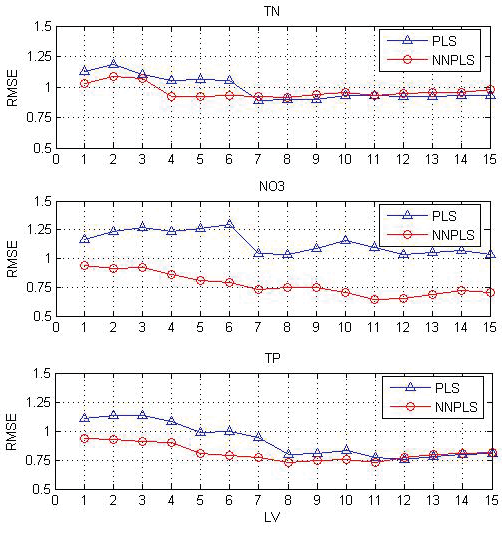

Figure 1. PLS and NNPLS prediction of TN, NO_3^- and TP. RRMSE as function of number of LV's.

3.2 Neural Net Partial Least Squares

Figure 1 shows the results for NNPLS modelling of TN. The optimal number of LV's is found to be 4 LV's. NNPLS thus captures the process behaviour in a lesser number of LV's. The relative RRMSE (relative root mean square error) for the 4-LV NNPLS model is however slightly higher (0.92) than the relative RRMSE for the 7-LV PLS model (0.89). In concordance with the latter, no improvement is seen in the prediction results (Figure 5).

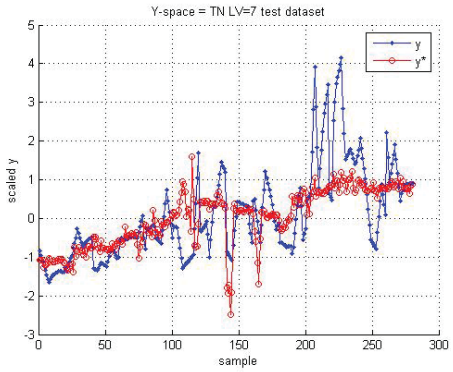

Figure 2. PLS prediction of TN. Original (y) and predicted (y*) data in the validation data set.

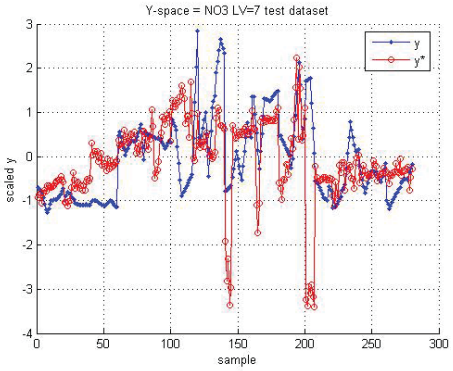

Figure 3. PLS prediction of NO_3^-. Original (y) and predicted (y*) data in the validation data set.

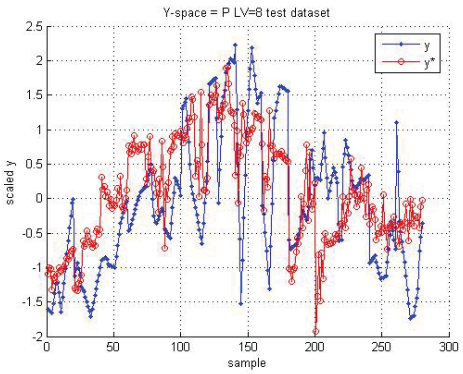

Figure 4. PLS prediction of P. Original (y) and predicted (y*) data in the validation data set.

For NO_3^- prediction, an 11-LV model was selected on the basis of the results shown in Figure 1. The resulting RRMSE value (0.64) is considerably lower compared to the RRMSE for the PLS model (1.04). This improvement is also reflected in Figure 6 when compared with Figure 3.

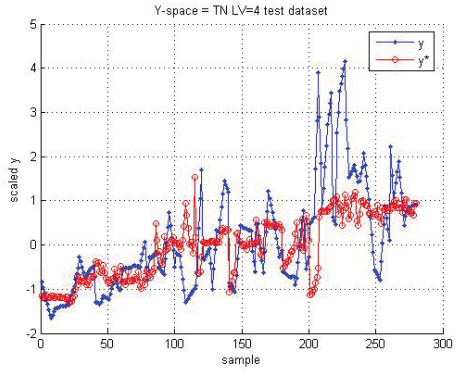

Figure 5. NNPLS prediction of TN. Original (y) and predicted (y*) data in the validation data set.

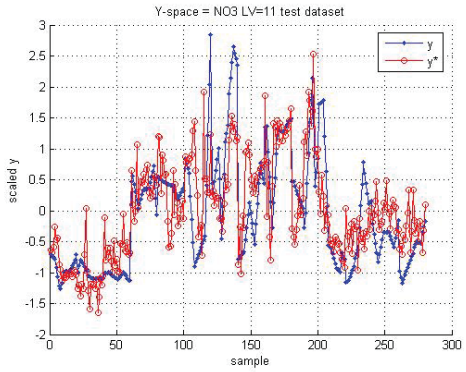

Figure 6. NNPLS prediction of NO_3^-. Original (y) and predicted (y*) data in the validation data set.

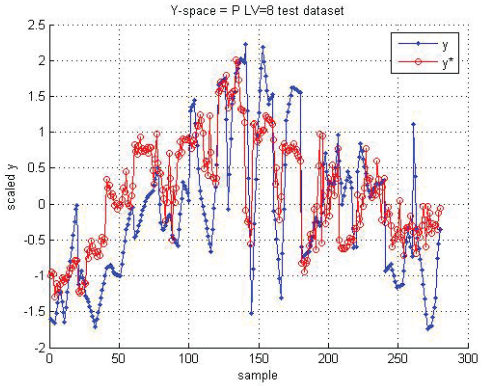

Figure 7. NNPLS prediction of P. Original (y) and predicted (y*) data in the validation data set.

Figure 7 presents the results regarding NNPLS regression of phosphorous (TP). The 8-LV model was selected. The RRMSE value (0.73) is lower than the RRMSE of the PLS model (0.80). In contrast to this reduction, improvement is harder to see when comparing NNPLS predictions (Figure 7) with PLS predictions (Figure 4).

3.3 Kernel Partial Least Squares

The RRMSE values for KPLS-based TN prediction are shown in Figure 8 as a function of the number of LV's, LV, and the kernel width, d. The minimum RRMSE (0.72) was found for 5 LV's and a kernel width of 196. Even though a lower RRMSE was obtained when comparing to the NNPLS model (0.92), the prediction results shown in Figure 9 hardly support the use of the KPLS model since the model is not able to track any but very slow dynamics.

Figure 10 shows the results for KPLS regression of NO_3^-. The best model (RRMSE = 0.84) is found for 2 LV's and a kernel width of 1510. The obtained RRMSE value is however higher than the RRMSE value for the NNPLS model (0.64). When the prediction results (Figure 11) are compared to the prediction of the NNPLS model, KPLS delivers a poor predictor, especially when considering the observed dynamics.

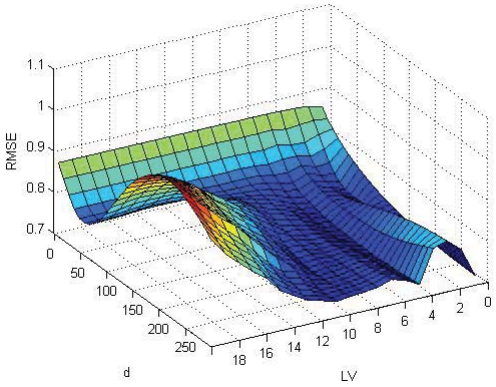

Figure 8. KPLS prediction of TN. RRMSE as a function of the number of LV's and the kernel width.

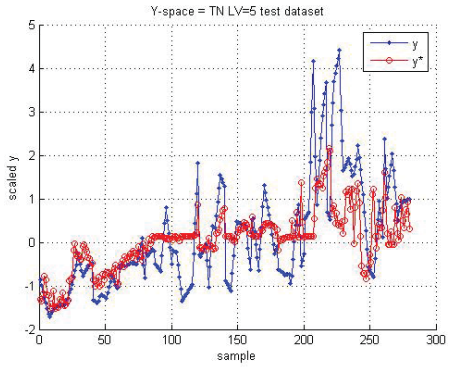

Figure 9. KPLS prediction of TN. Original (y) and predicted (y*) data in the validation data set.

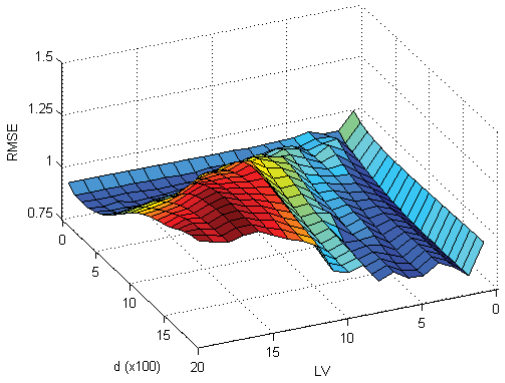

Figure 10. KPLS prediction of NO_3^-. RRMSE as a function of the number of LV's.

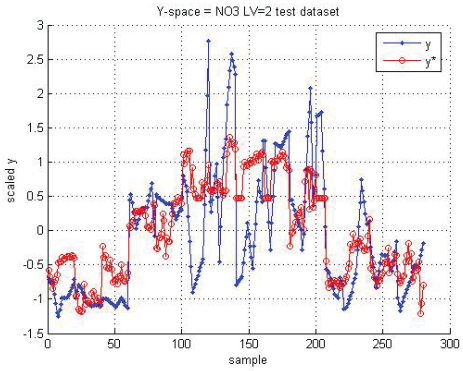

Figure 11. KPLS prediction of NO_3^-. Original (y) and predicted (y*) data in the validation data set.

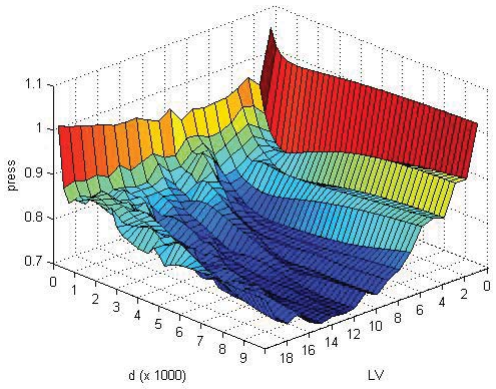

Figure 12. KPLS prediction of TP. RRMSE as a function of the number of LV's.

RRMSE values for the KPLS regression of TP are shown in Figure 11. The kernel width, d, was increased up to 10000, but no minimum was found within the evaluated range. At the border of this range, the 16 LV's and a kernel width of 10000 deliver the minimal RRMSE. Despite the lower RRMSE (0.71), when compared to the PLS (0.80) and NNPLS (0.73) model, the slight improvement seen in Figure 12 is questionable as KPLS needs twice as many LV's compared to the NNPLS model.

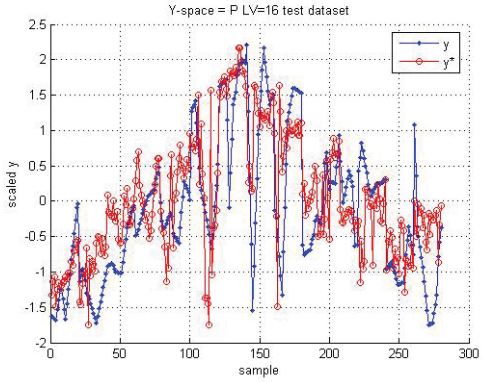

Figure 13. KPLS prediction of TP. Original (y) and predicted (y*) data in the validation data set.

4. DISCUSSION

Table 1 presents a (subjective) quality mark based on the discussed results, the minimal RRMSE values and the corresponding number of LV's found for each model type and response variable. With respect to NO_3^- and TP prediction, the NNPLS method delivered better results compared to PLS. For TN, the NNPLS model delivers the worst prediction, but in fact all models for TN are performing poorly. Despite the improved RRMSE values for TN and TP, the KPLS models are put in doubt due to their low short-term predictive power. Next to this, KPLS models come with large computational efforts, which questions their use in systems for on-line control. In our case, KPLS modelling typically demanded 10 to 20 times more time compared to the other models. Extra computational demands due to NNPLS modelling were negligible. As such, NNPLS is preferred.

Table 1. Summary of RRMSE values and selected number of LV's for all evaluated models

	output	PLS	NNPLS	KPLS
quality	TN	-	-	-
	NO3	-	++	-
	TP	+	+	+
RRMSE	TN	0.89	0.92	0.72
	NO3	1.04	0.64	0.84
	TP	0.80	0.73	0.71
LV's	TN	7	4	5
	NO3	7	11	2
	TP	8	8	16

More importantly, improvement of the resulting models may be obtained when accounting for the following hypotheses:

- The on-line data do not capture the TN-related processes completely, i.e. the process is not observable given the on-line process data.

- The observations were treated as independent observations. Ignoring auto-correlation in the data may induce a serious deterioration of the model quality.
- The dataset represents batches within a period longer than 14 months, exhibiting considerable changes in operation. A well-performing model generalising over the whole dataset may not be feasible.

Model prediction performance may be considerably improved if models are made locally in time. While they may counter the problem of autocorrelation, such an approach may also circumvent problematic modelling due to changing system behaviour. Next to that, accounting for possible autocorrelation in the data set, e.g. by means of ARX structuring, may result in considerable model improvement. This, however, will likely result in a considerable increase of computational requirements.

5. CONCLUSIONS

In this work, PLS, NNPLS and KPLS models were constructed for prediction of effluent quality variables (TN, NO_3^- and TP) on the basis of on-line process data (V, T, conductivity, DO, ORP and pH). It was shown that the NNPLS models deliver best results compared to the PLS models in the case of nitrate and total phosphorous. Less trust exists with the KPLS models. Despite these conclusions, it is suggested that improvement may be obtained when models are made locally in time and/or when potential autocorrelation is accounted for.

6. ACKNOWLEDGEMENTS

This work was supported by the Institute for Encouragement of Innovation by means of Science and Technology in Flanders (IWT) and the ERC program of MOST/KOSEF (R11-2003-006-01001-1) through the Advanced Environmental Biotechnology Research Center at POSTECH. Peter Vanrolleghem is Canada Research Chair in Water Quality Modelling.

7. REFERENCES

Geladi P., Kowalski, B.R., Partial least-squares regression: a tutorial. *Analytica Chimica Acta, 185,* 1-17, 1986.

Gurden S.P., Westerhuis, J.A., Bro, R. and Smilde, A.K., A comparison of multiway regression and scaling methods. *Chemometrics and Intelligent Laboratory Systems, 59,* 121-136, 2001.

Insel G., Sin G., Lee D.S., Nopens I. and Vanrolleghem P.A., A calibration methodology and model-based systems analysis for SBRs removing nutrients under limited aeration conditions. *Journal of Chemical Technology and Biotechnology., 81,* (in Press), 2006

Kourti T. and MacGregor J.F., Process analysis, monitoring and diagnosis, using multivariate projection methods, *Chemometrics and Intelligent Laboratory Systems, 28,* 3-21, 1995.

Lee D.S., Vanrolleghem P. and Park J.M., Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant, *Journal of Biotechnology, 115,* 317-328, 2005.

Mujunen, S.-P., Minkkinen, P., Teppola, P. and Wirkkale, R.-S., Modeling of activated sludge plants treatment efficiency with PLSR: a process analytical study, *Chemometrics and Intelligent Laboratory Systems, 41,* 83-94, 1998.

Nomikos, P., & MacGregor, J. F., Multi-way partial least squares in monitoring batch processes, *Chemometrics and Intelligent Laboratory Systems, 30,* 97-108, 1995.

Rosipal, R. and Trejo, L.J., Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space, *Journal of Machine Learning Research, 2,* 97-123, 2001.

Schölkopf, B., Mika, S., Burges, C. J. C., Knirsch, P., Müller, K.-R., Rätsch, G., and Smola, A. J., Input space versus feature space in kernel-based methods, *IEEE Transactions on Neural Networks, 10(5),* 1000-1016, 1999.

Sin G., Villez K. and Vanrolleghem P.A., Application of a model-based optimisation methodology for nutrient removing SBRs leads to falsification of the model, *Water Science and Technology, (In press),* 2006.

Teppola, P, Mujunen, S.-P. and Minkkinen P., Partial Least Squares modeling of an activated sludge plant: A case study, *Chemometrics and Intelligent Laboratory Systems, 38,* 197-208, 1997.

Wilderer, P.A., Irvine, R.L. and Goronszy, M.C., Sequencing Batch Reactor technology, *Scientific and Technical report Series No. 10, IWA publishing,* pp100, 2001.

Wold, S., Kettaneh, N. and Skagerberg, B., Non-linear PLS modelling, *Chemometrics and Intelligent Laboratory Systems, 7,* 53-65, 1989.

Wold, S., Kettaneh, N., Fridén, H., and Holmberg A., Modelling and diagnostics of batch processes and analogous kinetic experiments, *Chemometrics and Intelligent Laboratory Systems, 44,* 331-340, 1998.