Indonesian Journal of Science & Technology

Journal homepage: http://ejournal.upi.edu/index.php/ijost/

# Sludge Dewatering Process Control using Principal Component Analysis (PCA) and Partial Least Square (PLS)

*Grace Pebriyanti [1]*, Renjie Zhu[2], and Adelhard Beni Rehiara[1,3]*

[1] Engineering Department, Universitas Papua, Jl. Gunung Salju, Manokwari 98314, Indonesia
[2] Junior Engineering, Tebodin Engineering & Construction, Rotterdam, The Netherlands
[3] Department of Cybernetics, Graduate School of Engineering, Hiroshima University, Japan

*Corresponding author: Email: grace.p.tambun@gmail.com

A B S T R A C T S

The process control in the sludge dewatering process is to minimalize the water volume in the sludge. However, management of this process control is difficult because of its multi-variables, nonlinearity and long delay. In this paper, a control approach based on the principal component analysis (PCA) is presented. A PCA model, which incorporates time lagged variables is used. The control objective is expressed in the score space of this PCA model. A controller is designed in the model predictive control framework, and it is used to control the equivalent score space representation of the process. The score predictive model for the model predictive control algorithm is built using a partial least squares (PLS). The process control system with PLS was simulated on Matlab and the graphs showed good accuracy and stability.



A R T I C L E I N F O

*Article History:*
*Received 2 Jan 2016*
*Revised 18 Feb 2016*
*Accepted 23 Feb 2016*
*Available online 29 Mar 2016*

*Keywords:*
Process control
Principal component analysis
Partial least squares
Sludge dewatering

## 1. INTRODUCTION

Sludge a dewatering process is a nonlinear process with a long delay. Producing the sludge with high dryness content is an important process control objective. The sufficient dryness content of the sludge dewatering process is often to be not satisfactory due to the technical limitations of the available centrifuge used in the on-line sludge dewatering process. Moreover, there are many parameters that have to be observed such as flow rate of the sludge, flow rate of the polymer, dry solids in supply, temperature, solid content (centrate), sludge cake dryness, differential speed of the centrifuge, and the screw torque of the centrifuge.

To solve this problem, many approaches have been studied from many researches which have similar basic principle of the process. Thyagarajan *et al.* presents an application of artificial neural network (ANN) technique to develop a model representing the nonlinear drying process. (Thyagarajan *et al.*, 1997) Trelea *et al.* presented a simple nonlinear process predictive optimal control algorithm for on-line control of a batch drying process. (Trelea *et al.*, 1997) Suykens *et al.* presented weighted least square support vector machine. (Suykens *et al.*, 2002) Turovskiy *et al.* presented the waste water sludge processing. (Turovskiy *et al.*, 2006)

In this paper, we a proposed a new approach for a process control conditon for using when accurate moisture measurements are not available on-line or have a long time delays. The key aspects of this controller are as follows:

(1) A PCA model that uses time – lagged data. The scores calculated from this model are fed as inputs to a score predictive model which is developed using a partial least square (PLS).

(2) The predicted scores are used as key indicators of the process performance based on the assumption of an implicit correlation between available measurements used to calculate the scores and the cake dryness variables.

(3) The control objective is that used to maintain the predicted score variables within a certain acceptable region defined from historical data. This controller is developed and implemented on top of an existing plant-wide conventional PID control system. Manipulated variables for the proposed controller are selected in the set points of existing control loops. Other issues (i.e. maintaining the correlation structure of the input variables when implementing the controller) we also discussed. In the following sections, the methods used in this paper are briefly reviewed, and the design of the control system is presented. The final section draws conclusions regarding this study.

## 2. METHOD

### 2.1. Principal Component Analysis (PCA)

Originally principal component analysis or PCA was developed by Pearson in 1901. PCA is a method for analyzing the covariance of a data set of plant variables. The approach transforms a matrix containing measurements from n process variables, **X**, into a matrix of mutually uncorrelated variables. These variables (called principal components (PC)) are transformed from the original data into a new basis defined by a set of orthogonal loading vectors, $\mathbf{p_k}$. The individual values of the principal components are called scores. This transformation is determined as follows:

$$\boldsymbol{X} = \sum_{k=1}^{np<n} \boldsymbol{t_k}\, \boldsymbol{p_k^T} + \boldsymbol{E} \qquad (1)$$

DOI: http://dx.doi.org/10.17509/ijost.v1i1
p- ISSN 2528-1410 e- ISSN 2527-8045

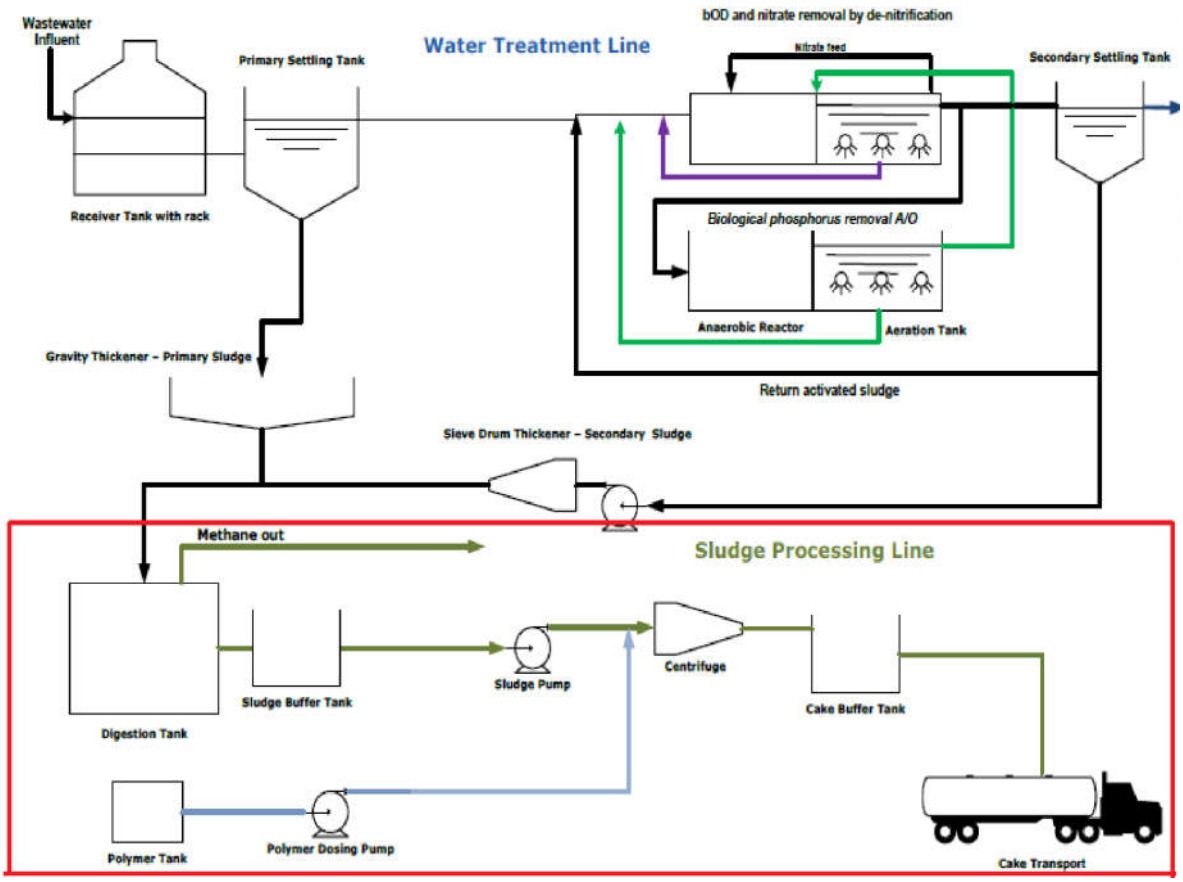


**Figure 1.** The block scheme of the Sludge Dewatering Process.

The loadings are defined here as being orthonormal, and so they become the eigenvectors of the data covariance matrix, $\mathbf{X}^T\mathbf{X}$. The $\mathbf{t}_k$ and $\mathbf{p}_k$ pairs are ordered so that the first pair captures the largest amount of variation in the data and the last pair captures the least. In this way, it is generally found that a small number of PCs (**np**) can account for much of the power in the covariance matrix. The remaining power constitutes the error term **E**. When Eq.(1) is applied to a single vector of new process measurements, $\mathbf{X}^T$, the resulting term **E** is called the prediction error.

### 2.2. Partial Least Square (PLS)

PLS regression is a recent technique that generalizes and combines features from principal component analysis and multiple regressions. It is particularly useful when we need to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors). It originated in the social sciences (specifically economy) but became popular in chemometrics (i.e., computational chemistry) due in part to Herman's son Wvante, (Geladi & Kowalski, 1986)) and in sensory evaluation (Martens & Naes, 1989). But, PLS regression is also becoming a tool of choice in the social sciences as a multivariate technique for non-experimental and experimental data alike (e.g., neuroimaging (Mcintosh *et al.*, 1996)). It was first presented as an algorithm akin to the power method (used for computing eigenvectors) but was rapidly interpreted in a statistical framework.

The goal of PLS regression is to predict **Y** from **X** and to describe their common structure. To prerequisite notions and notations, the **I** observations described by **K** dependent variables are stored in the **I** × **K** matrix denoted **Y.** The values of **J** predictors collected on these **I** observations are

DOI: http://dx.doi.org/10.17509/ijost.v1i1
p-ISSN 2528-1410 e-ISSN 2527-8045

collected in the **I** × **J** matrix denoted **X**. When **Y** is a vector and **X** is full rank, this goal could be accomplished using ORDINARY MULTIPLE REGRESSION. When the number of predictors large compared to the number of observations, **X** is likely to be singular, and the regression approach is no longer feasible because of MULTICOLINEARITY. Several approaches have been developed to cope with this problem. One approach is to eliminate some predictors (e.g., using step-wise methods) of another one, called principal component regression. This is to perform PRINCIPAL COMPONENT ANALYSIS (PCA) of the **X** matrix and then is used the principal components of **X** as regressors on **Y**. The orthogonality of the principal components eliminates the multicolinearity problem. But, the problem of choosing an optimum subset of predictors remains. A possible strategy is to keep only a few of the first components. But they are chosen to explain **X** rather than **Y.** Indeed, nothing guarantees that the principal components (which "explain" **X**) are relevant for **Y**.

In contrast to the above exploration, PLS regression finds components from **X** that is also relevant for **Y**. Specifically, PLS regression searches for a set of components (called latent vectors). It that performs a simultaneous decomposition of **X** and **Y** with the constraint, in which these components explain as much as possible of the covariance between **X** and **Y**. This step generalizes PCA. It is followed by a regression step where the decomposition of **X** is used to predict **Y**.

The approach is worked by selecting factors of cause variables in a sequence which successively maximize the explained covariance between the cause and effect variable. A matrix of cause data given, **X**, and the effect data, **Y**, a factor of the cause data, $\mathbf{t_1}$, and the effect data, $\mathbf{u_1}$, is evaluated by

$$X = \sum_{h=1}^{np<nx} t_h p_h^T + E, \qquad Y = \sum_{h=1}^{np<nx} t_h q_h^T + F$$

where **E** and **F** are the residual matrices and **np** is the number of inner components that are used in the model.

These equations are referred as the outer relationships. The vectors $\mathbf{t_h}$ are mutually orthogonal. These vectors and $\mathbf{u_h}$ are selected so as to maximize the covariance between each pair, ($\mathbf{t_h}$, $\mathbf{u_h}$). Linear regression is performed between $\mathbf{t_h}$ and **uh**, to produce the inner relationship the correlation can be written as:

$$u_h = b_h t_h + \varepsilon_h \qquad (2)$$

where $\mathbf{b_h}$ is a regression coefficient and $\varepsilon_h$ refers to the prediction error. The PLS method provides the potential for a regularized model through selecting an appropriate number of latent variables, $\mathbf{u_h}$ in the model (**np**) (Geladi & Kowalski, 1986). More concepts of least square estimation for the static systems identification is explained by Stigter. (Stigter, 2011)

### 2.3. Design of a process controller

Since the variables of sludge dewatering process are almost never independent one to another, the true dimension of the space can be described as the process moves is usually very much smaller than the number of measured variables. As a result, many of the measured variables move together because of a few underlying fundamental events indeed, affects this the entire process and leads to the latent variable approach used. The sludge dewatering process scheme is shown in **Figure 2**.

In order to design a process control of the sludge dewatering process, the process definition has to be determined. The control goals are to determine the static model of the sludge dewatering process to make this process can run at its absolute best condition and to optimize the parameters that are running at various condition (i.e. the sludge feed rate, the polymer dosing, the sludge cake dryness, and the low concentrate). Increasing the dry solid recovery means reducing the recycle load as well as saving the energy or money. The

DOI: http://dx.doi.org/10.17509/ijost.v1i1
p- ISSN 2528-1410 e- ISSN 2527-8045

sludge dewatering process parameters are defined on **Table 1**.

Three abbreviations used in the process controller are (i) MV (Manipulated Variable) for a process input that can be independently set by the controller, (ii) DV (Disturbance Variable) for a measurable process input that affects the process outputs, (but cannot be set by the controller), and (iii)CV (Controlled Variable) for a process output controlled by the control.

## 3. RESULTS AND DISCUSSION

Assume that there is a linear relationship between parameters. However, since the dewatering sludge process is a dynamic process, thus non-linearity should take into account. If there is no non-linearity, then a good prediction from the PLS will be sufficient as long as the chosen parameters are correct. These are the assumptions before doing the modeling in order to help us making a good predictor.

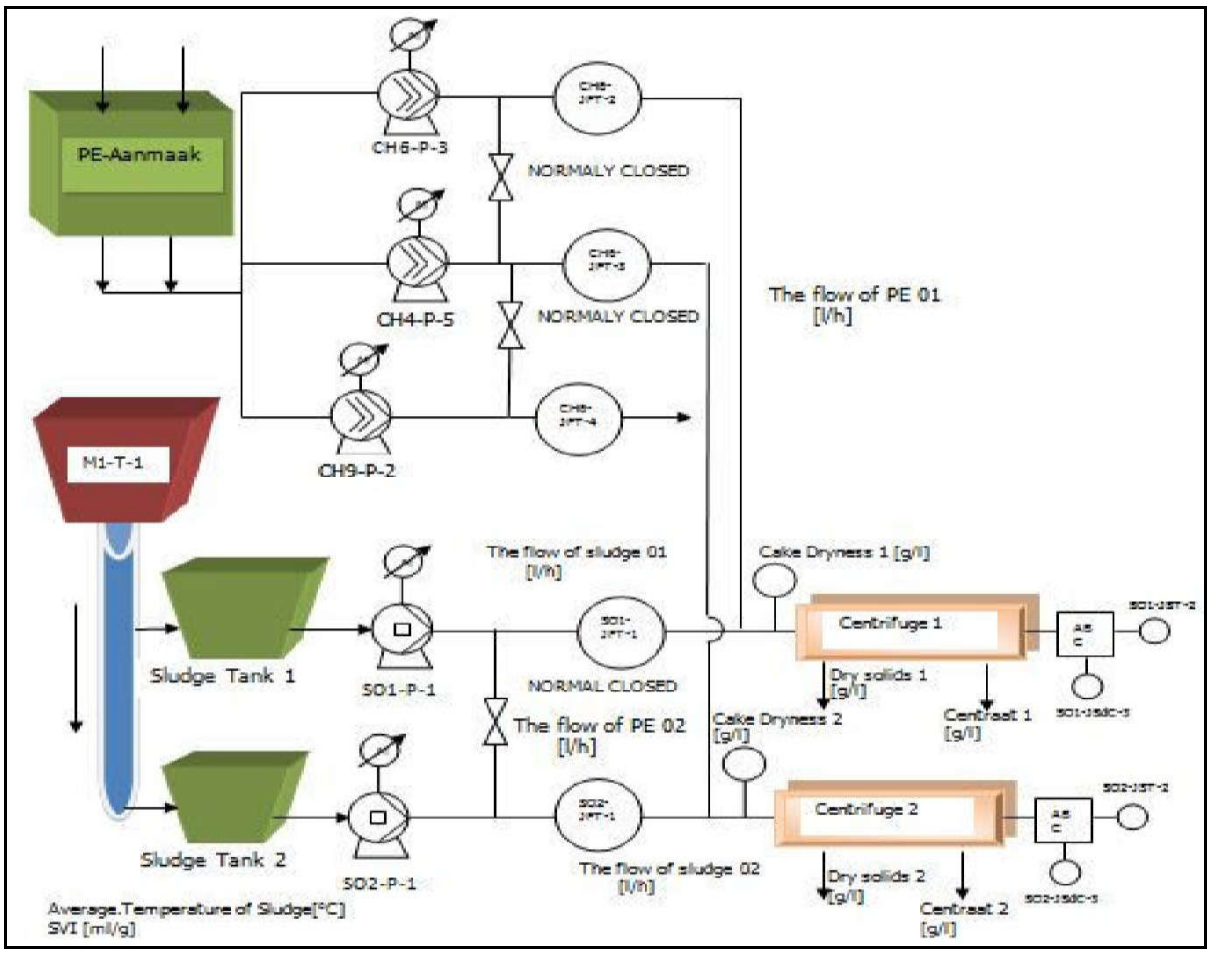


**Figure 2** The scheme of the sludge dewatering process.

DOI: http://dx.doi.org/10.17509/ijost.v1i1
p-ISSN 2528-1410 e-ISSN 2527-8045

**Table 1.** The process parameters for the process controller

| Parameter Name | MC/CV/DV | Accuracy | Response Time | Data Availability |
|---|---|---|---|---|
| Flow rate of sludge | DV: flow rate [$m^3/s$] | Flow sensor | Min/Sec | Available |
| Flow rate of Polymer | CV: flow rate [$m^3/s$] | Flow sensor (Control motor) | Min/Day | Available |
| Dry solids in supply | MV: ratio (%) | Laboratory | Day/Week | Available |
| SVI | DV: ratio (%) | Laboratory | Day/Week | Available |
| Temperature | DV: temperature changing ($^oC$) | Temperature sensor | Day/Week | Available |
| Concentrate | MV: solids content [g/l] | Laboratory | Day/Week | Available |
| Cake Dryness | MV: ratio (%) | Laboratory | Day/Week | Available |
| Differential speed and screw torque | CV: bowl speed [m/s] screw torque [N.m] | Speed sensor and torque sensor | Min/Sec | Not Available |

The static model for determining the relationship between the parameters is shown on **Table 2** based on four input parameters. From the above table, there is a correlation between parameters because of several reasons: (1) the original data is so poor, and (2) there is a nonlinearity factor. Thus, we need to put more parameters such as torque, differential speed, seasonal situation, sludge characteristic, concentrate, temperature, and SVI. On this research, only temperature, SVI, and concentrate are added because of technical limitation to get the data of additional parameters. By analyzing 5 input parameters and 1 output parameter, the static model is shown on **Table 3**.

From **Table 3**, the most important parameters which have a strong relationship are found. The important parameters are dry solids in supply, concentrate, temperature, and cake dryness. **Figures 3** and **4** show the principal component analysis related to the variance of 5 inputs variables and the Pareto function explaining 2/3 of the total variance respectively.

DOI: http://dx.doi.org/10.17509/ijost.v1i1
p- ISSN 2528-1410 e- ISSN 2527-8045

**Table 2.** The static model of 4 inputs parameters

| Corelation coefficient | Flowrate Sludge2 | Flowrate PE2 | Dry solids in Supply | Cake Dryness |
|---|---|---|---|---|
| Flowrate Sludge2 | 1 | 0.045 | -0.118 | -0.079 |
| Flowrate PE2 | 0.045 | 1 | 0.166 | -0.164 |
| Dry solids in Supply | -0.118 | 0.166 | 1 | 0.379 |
| Cake Dryness | -0.079 | -0.164 | 0.379 | 1 |

**Table 3.** The static model of 5 input parameters and 1 output parameter

| Correlation coefficient | Flow-rate Sludge 2 | Flow-rate PE2 | Dry Solids in Supply | Con-centrat | Tem-pera-ture | Cake dry-ness |
|---|---|---|---|---|---|---|
| Flowrate Sludge2 | 1 | 0.18 | -0.26 | -0.15 | -0.11 | -0.1 |
| Flowrate PE2 | 0.18 | 1 | 0.17 | -0.004 | 0.08 | 0.06 |
| Dry Solids in Supply | -0.26 | 0.17 | 1 | -0.199 | 0.66 | 0.56 |
| Concentrate | -0.15 | -0.004 | -0.199 | 1 | -0.35 | -0.47 |
| Temperature | -0.11 | 0.08 | 0.66 | -0.35 | 1 | 0.67 |
| Cake dryness | -0.1 | 0.06 | 0.56 | -0.47 | 0.67 | 1 |

Choosing the number of components in the PLS model is a critical step. The graft in **Figure 4** gives a rough indication, showing around 40% of the variance explained by the first component, with as many as the second and third components making significant contributions reaching 80% of the variance. By using the partial least square method, the data cloud is built. The data cloud describing the data distribution according to the relationship between each parameter is presented on **Figure 5**. As a remark from this **Figure**, Flow rate Sludge2 and Flow rate PE2 have a similar character as well as Temperature and Dry solids in supply. These parameters are connected strongly as the cause variables for the output that is Concentrate as the effect variable. So that, these parameters are observed on partial least square modeling that can be seen on **Figure 6** showing a reasonable correlation between fitted and observed responses, and this is confirmed by the $R^2$ statistic (that is R squared = 0.6788).

DOI: http://dx.doi.org/10.17509/ijost.v1i1
p-ISSN 2528-1410 e-ISSN 2527-8045

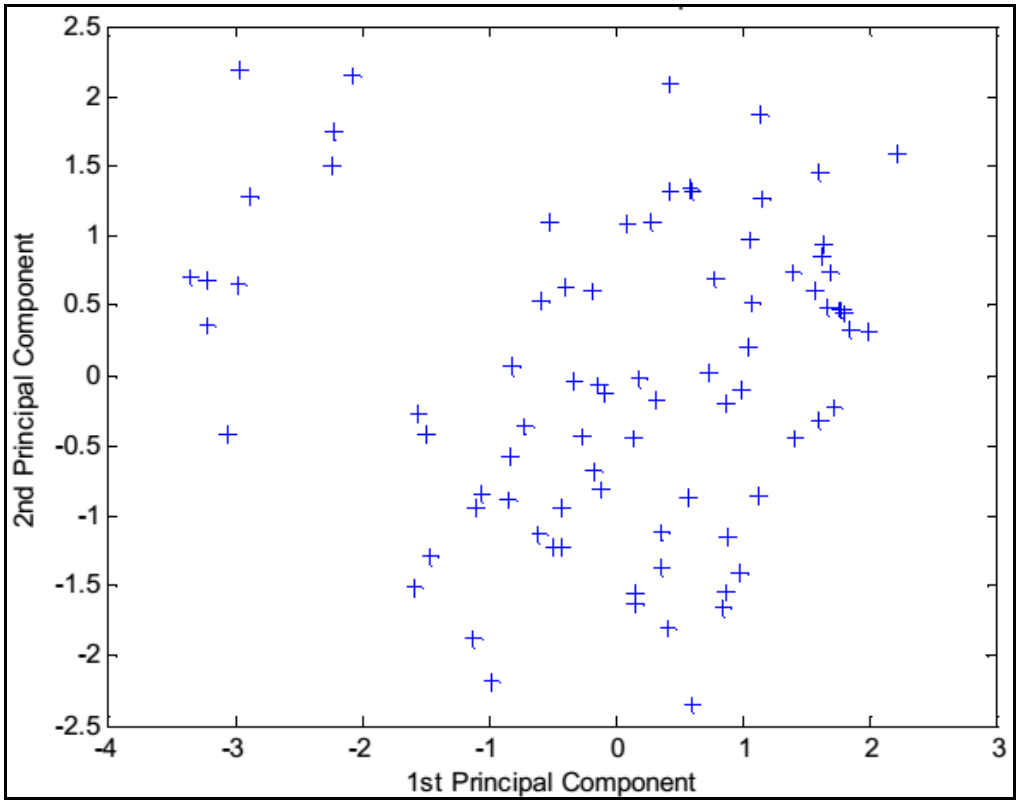


**Figure 3.** The principle component analysis relating to the five input variables

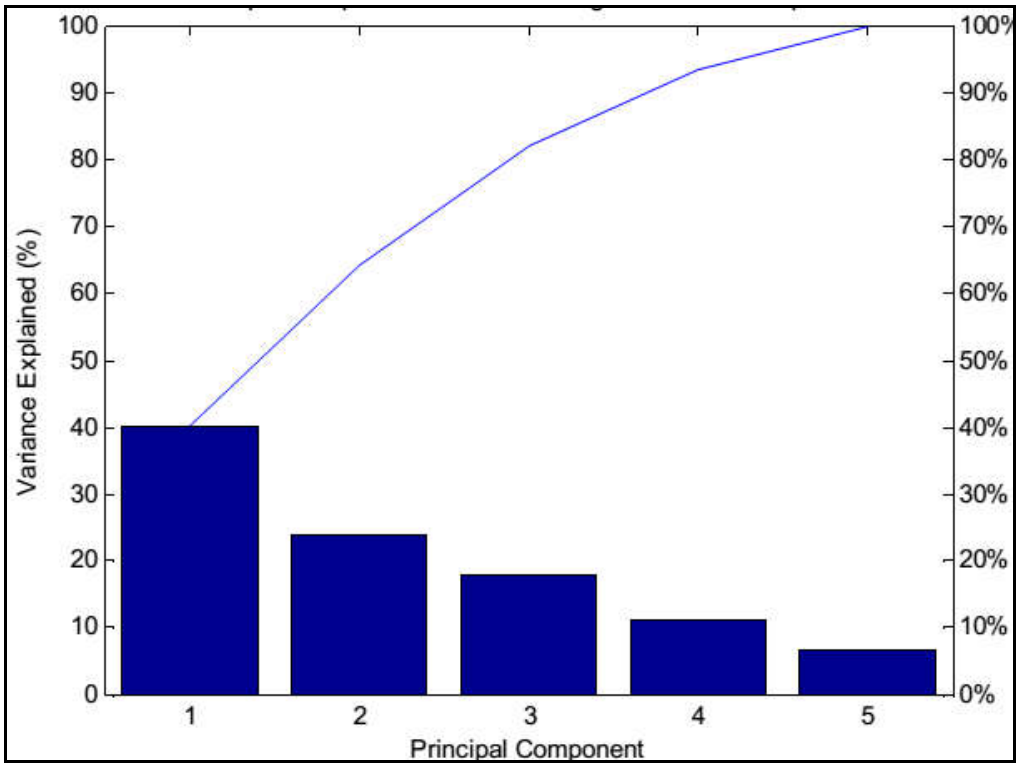


**Figure 4.** Principle components and percentage of variance

DOI: http://dx.doi.org/10.17509/ijost.v1i1
p- ISSN 2528-1410 e- ISSN 2527-8045

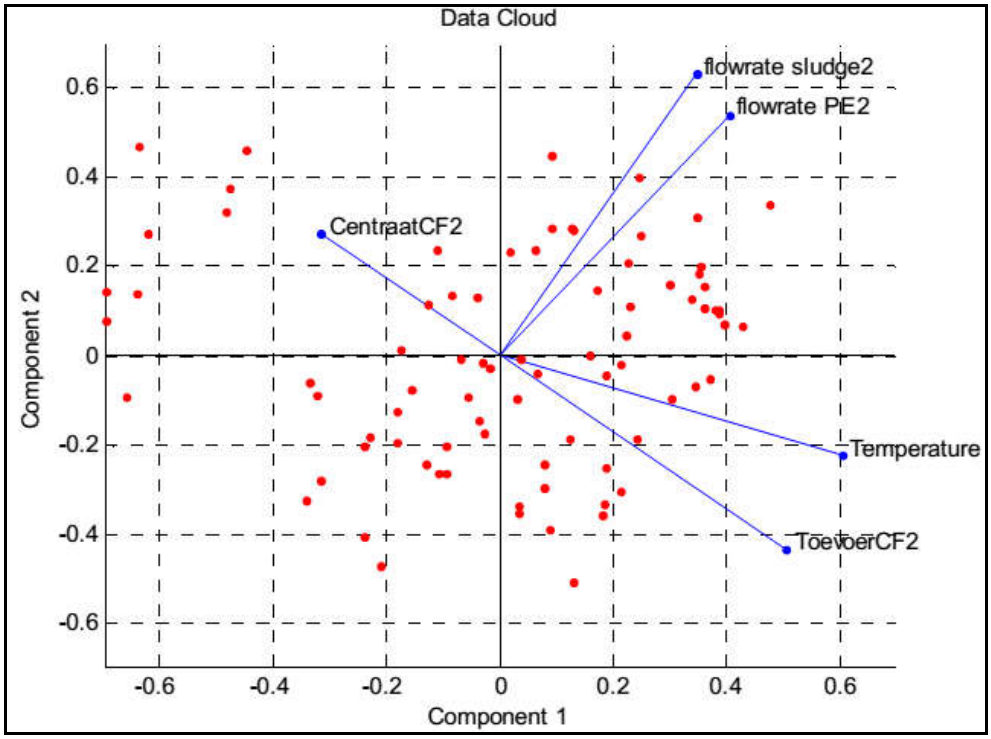


**Figure 5.** The data cloud of five input parameters.

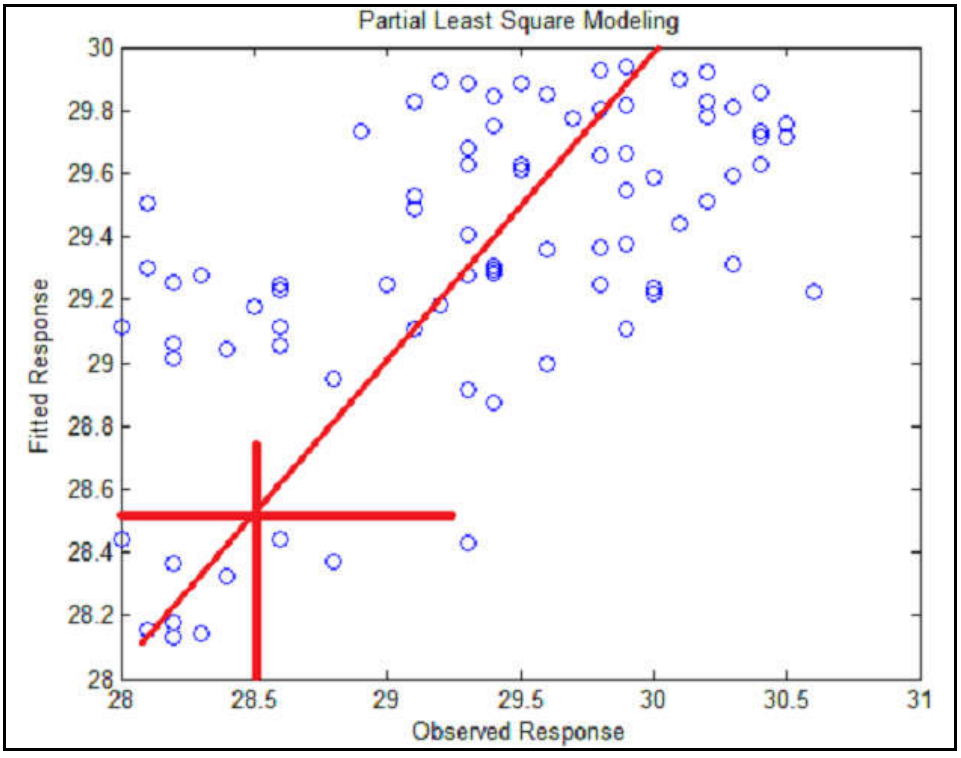


**Figure 6.** The partial least square modeling.

After doing the partial least square modeling, a plot of the weight of five predictors in each three components shows that two of the components (the blue and

DOI: http://dx.doi.org/10.17509/ijost.v1i1
p-ISSN 2528-1410 e-ISSN 2527-8045

green lines) explain the majority of the variance in the cause of the process. It can be seen from **Figure 7** that the two lines stick to close with a little difference around 0.05 in weight. Thus, a plot of the mean-squared errors can be done.

From **Figure 8**, it can be seen that the blue line is the MSE Predictors and the green line is the MSE Response. The plot of the Mean-Squared Errors (MSE) suggests that as few as two components may provide an adequate model. The error for the response is lower than the predictor.

From **Figure 9** and **Figure 10**, some remarks have been found. First, for the Predicted Model versus the Training Data, the error square value ($R^2$) is 0.4087 and the mean-squared error (MSE) is 1.0398e-015; for the Predicted Model versus the Validation Data, the error square value ($R^2$) is 0.5709 and the mean-squared error (MSE) is 0.0402. Thus, the Predicted Model fulfills the research objective sufficiently with a small error.

Secondly, there is the same trend for all data samples. The trend is that the Modeled is always trying to follow the training data. Secondly, for the Data Training on higher points of Dry Solids in centrifuge effluent, the Model cannot fit these points. The same response also happens when the Data Training on lower points. This phenomenon explains that there is another factor effecting on this process. This factor does not take into account on this paper. Probably, this happens because of a nonlinearity factor, which can not deal robustly with PLS method.

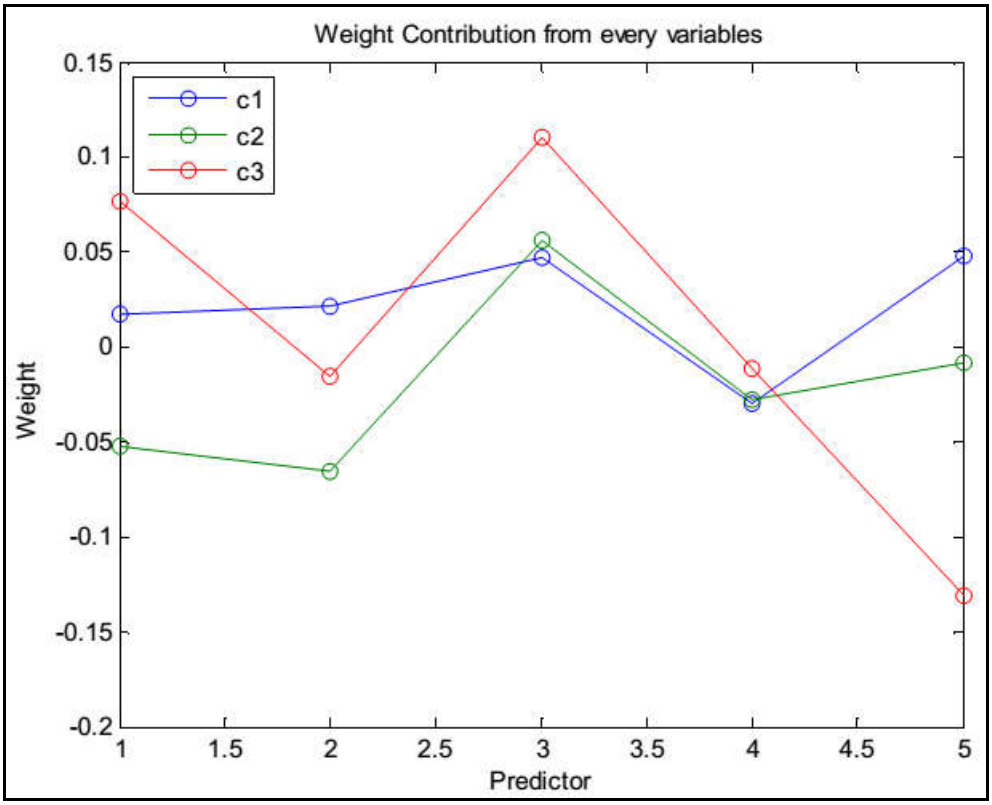


**Figure 7.** The weight contribution of variables.

DOI: http://dx.doi.org/10.17509/ijost.v1i1
p- ISSN 2528-1410 e- ISSN 2527-8045

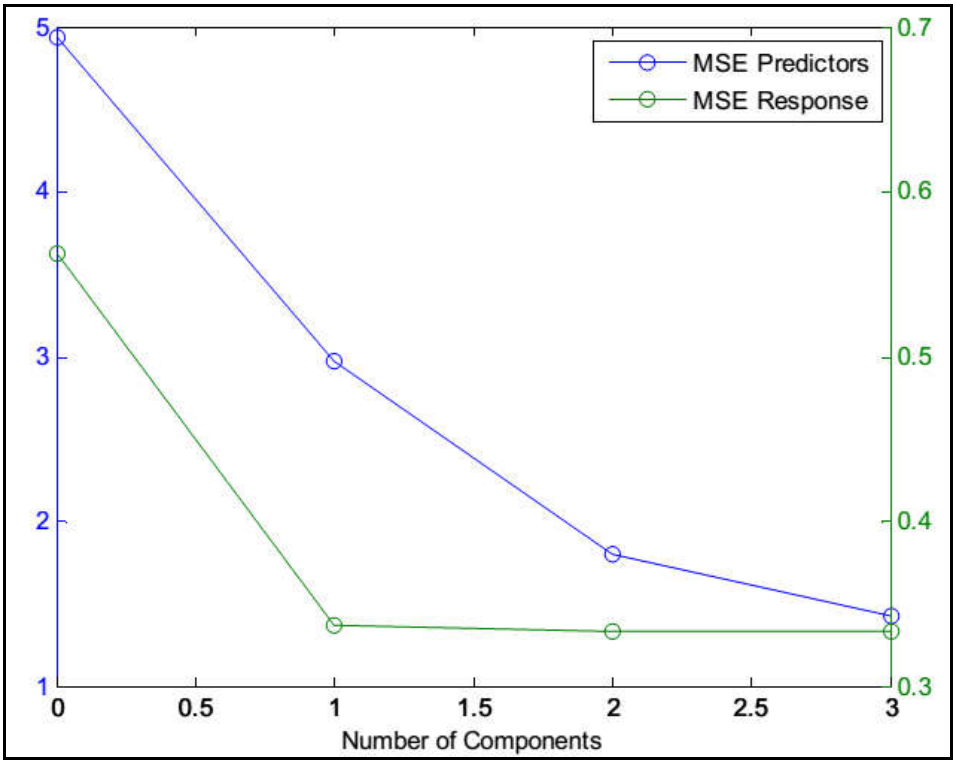


**Figure 8.** The graph of MSE predictors and MSE response.

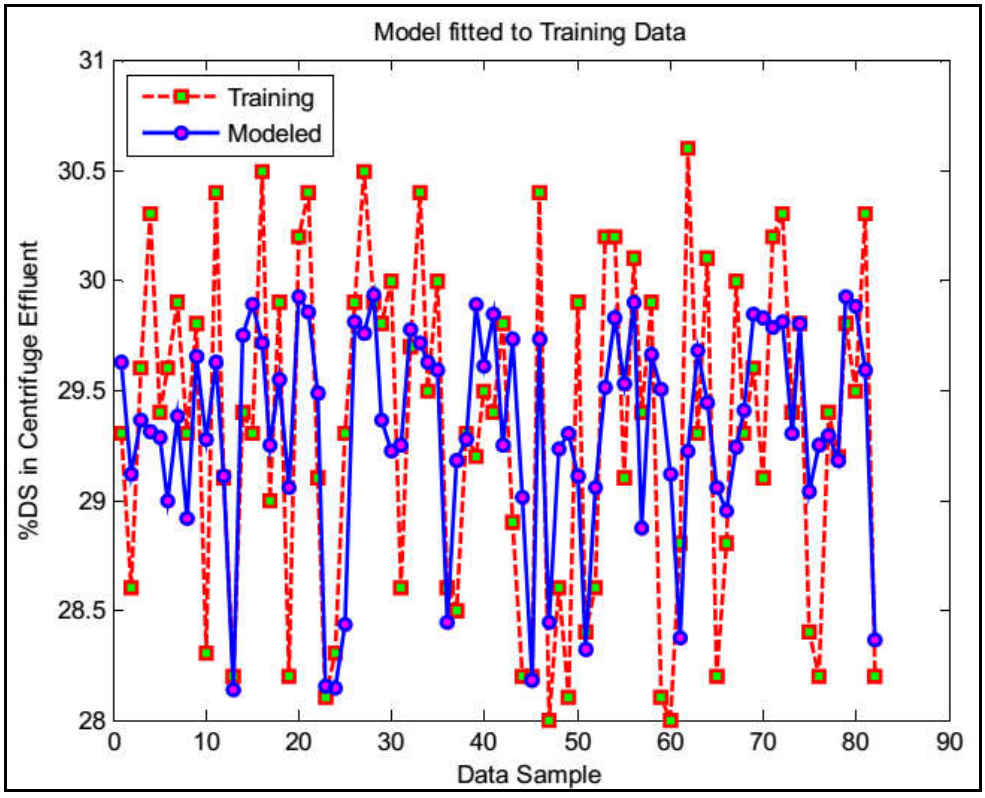


**Figure 9.** Predicted model versus the training data.

DOI: http://dx.doi.org/10.17509/ijost.v1i1
p-ISSN 2528-1410 e-ISSN 2527-8045

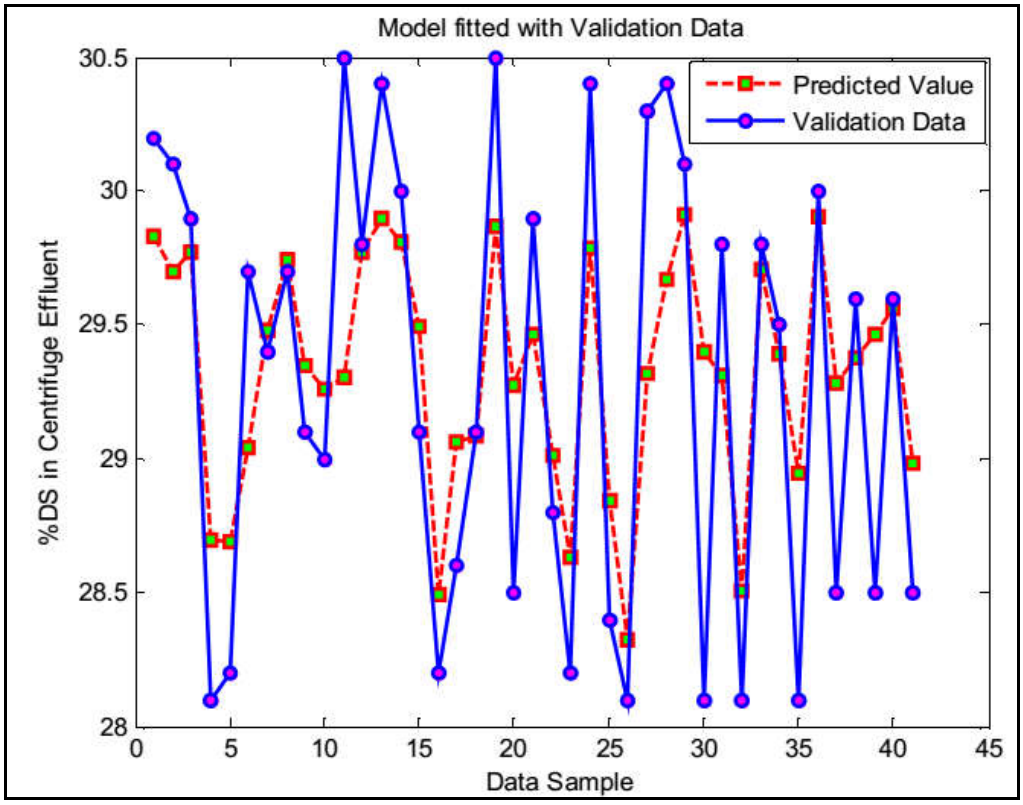


**Figure 10.** The Scheme of the sludge dewatering process.

## 4. CONCLUSIONS

Based on the above results, the conclusions are derived. Firstly, the principal component analysis (PCA) and partial least square (PLS) method have been successfully applied to real application performance modeling considered as a static modeling. This study applied PCA to analyze covariance of the data set of the sludge dewatering process. The parameters have the strong correlation to the result of the sludge dewatering process. The parameter are Dry solids in supply, Cake dryness, Temperature, and Concentrate. PLS method is applied to predict the model of the effect data set from the cause data set and to describe their common structure. The fixed size PLS with two components may provide a good model. The simulation results using the training data indicated that the model can fit the data in the range around 28% – 30% dry solid in centrifuge effluent. The same result has been determined on the simulation of the predicted model with the validation data. Thus, the simulation results presented that using PCA and PLS method are good for static system identification. Partial least square estimation is a reasonable way to estimate the unknowns from given data by calculating the prediction errors or residuals which are small. Lastly, use another approach such as LSSVM (Least Square Support Vector Machines) to improve the accuracy and robustness.

## 5. ACKNOWLEDGMENTS

This research would not have been possible without the support of the Control Systems Engineering Master Programme of HAN University of Applied Sciences as well as the Waste Water Treatment Plant (WWTP) Niewgraaf of the Waterboard Rijn en IJssel in the Netherlands. The authors

would also like to thank for the NESO – Nuffic Programme for financially supporting this research.

## 6. AUTHOR'S NOTES

The author(s) declare(s) that there is no conflict of interest regarding the publication of this article. Authors confirmed that the data and the paper are free of plagiarism.

## 7. REFERENCES

Geladi, P., and Kowlaski, B. (1986). Partial least square regression: A tutorial. *Analytica chemica acta*, 35, 1-17.

Martens, H., and Naes, T. (1989). *Multivariate Calibration*, London: Wiley.

McIntosh, A. R., Bookstein, F. L., Haxby, J. V., and Grady, C. L. (1996). Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage*, 3(3), 143-157.

Stigter, J. D. (2011). System Identification: An introduction. In MCSE programme HAN University. Netherlands.

Suykens, J. A., De Brabanter, J., Lukas, L., and Vandewalle, J. (2002). Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 48(1), 85-105.

Thyagarajan, T., Panda, R. C., Shanmugan, J., Rao, V. P. G., and Ponnavaikko, M. (1997). Development of ANN model for non-linear drying process. *Drying technology*, 15(10), 2527-2540.

Trelea, I. C., Trystram, G., and Courtois, F. (1997). Optimal constrained non-linear control of batch processes: application to corn drying. *Journal of food engineering*, 31(4), 403-421.

Turovskiy, I. S., and Mathai, P. K. (2006). *Wastewater Sludge Processing*. John wiley and sons.