PROCESS STATES AND THEIR SUBMODELS USING SELF-ORGANIZING MAPS IN AN ACTIVATED SLUDGE TREATMENT PLANT

Mikko Heikkinen[1], Tomi Heikkinen[2], Yrjö Hiltunen[1]

[1]University of Kuopio, Department of Environmental Sciences,
P.O. Box 1627, FIN-70211 Kuopio, Finland
[2]UPM-Kymmene, Wisaforest, Pietarsaari,
WIF Support Center, Wisaforest, Luodontie 149,
FIN-68601 Pietarsaari, Finland

mikko.heikkinen@uku.fi (Mikko Heikkinen)

Abstract

This paper presents an overview of an analysis method based on Self-Organizing Maps (SOM) which was applied to an activated sludge process. The aim of the study was to determine whether the neural network modelling method could be a useful and time-saving way to analyze this kind of process data. The used analysis procedure went as follows. At first, the process data is modeled using the SOM algorithm. Next, the reference vectors of the map were classified by K-means algorithm into six clusters, which represented different states of the process. At the final stage, the reference vectors of the clusters were used as sub-models to indicate variable dependencies in different clusters. The results show that the method presented here can be a good way to analyze this type of process data.

1 Introduction

Today, environmental regulations set challenges for controlling the industrial emissions, while production will be increased. Although, the emission levels are nowadays rather low compared e.g. to the 1970.

An activated sludge treatment process is a biological method for treating waste water. Typically, it is based on three main stages; pre-sedimentation, aeration and secondary clarifying. This is a common treatment system in the pulp and paper industry.

Despite the fact that the behavior of activated sludge is basically known, long lags and the complex character of biological activity set a challenge for controlling the process. Malfunctions are possible e.g. concentrations of nutrients in effluents may grow and metabolism rates may vary. Generally, the quality of effluent and circumstances in an aeration basis has directly an effect on the quality of the biological sludge.

One of the most remarkable factors of the treatment process is sludge settling properties. For example, if the lack of oxygen grows, filamentous sludge leads into poor settling properties. The treatment process could cause trouble, if the sludge does not settle at the bottom of the secondary clarifier. In addition, there are a lot of well-known reasons, which affect the filamentous sludge. Unfortunately, many of these factors concerning sludge settling properties are difficult to measure and detect in practice.

Chemical Oxygen Demand (COD), like Biological Oxygen Demand (BOD), is one of the most important factors for monitoring and control of an activated sludge treatment process. In addition, pulp mills have limits for emission, such as COD, set by authorities. Therefore predictive modelling could bring useful information for the behavior of the treatment process. The prospective information can help process control to react earlier especially if the trend of emissions is increasing.

Archived process data is an important resource for the knowledge management of the process and it can be used e.g. for the optimization of the process and the classification of the process states. Many studies have shown the importance of data-based modelling methods such as neural networks in the field of an industrial process [1-8]. However, a data centric approach in process analyses can be difficult due to unknown lags and differences in the character of variables. Additionally, missing or erroneous data

complicate the modeling of large data sets, because most data processing applications require complete data matrices before analysis [9].

Here, we apply Self-Organizing Maps (SOM) to the analysis of an activated sludge treatment process.

2 The process and data

The production capacity of UPM-Kymmene Wisaforest pulp mill is 800 000 tons of pulp annually. The activated sludge treatment plant of the pulp mill treats an average 1000 l/s of waste water. The treatment plant has the following phases (see Figure 1); (i) pre-sedimentation and equilibration, (ii) aeration and (iii) secondary clarifying.

The raw data (on-line data and laboratory measurements) was extracted from databases of the pulp mill. Variable selection was made by a process expert. The resolution of the complete data have been set for one day and the size of the data matrix was 29 x 1378 (29 variables in columns, 1378 rows); i.e. 4 years. Variables used in modelling are presented in Table 1.

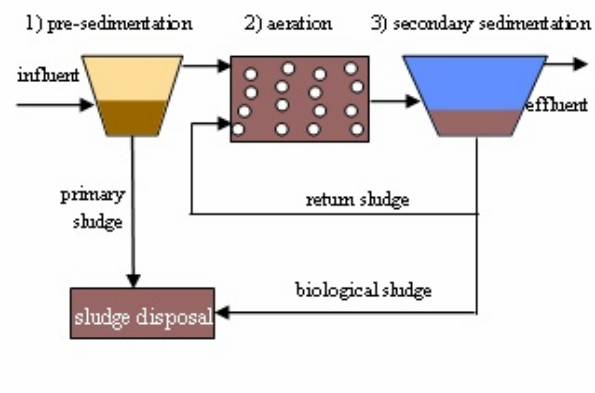

Fig. 1 A simplified diagram of the activated sludge treatment process with three main stages; 1) pre-sedimentation, 2) aeration and 3) secondary sedimentation.

3 Methods

3.1 Pre-processing the data

3.1.1 The missing data

Two kinds of missing data were observed in on-line measurements and in measurements made in the laboratory. The missing data in the first case has been caused by measurement errors and breakdowns of the measurement equipment. The missing data in laboratory measurements have been caused by low sampling frequencies.

The eventual imputation of missing data was then achieved using simple LI-algorithm, which fills the gaps in a table by drawing a straight line between two neighboring values and returning the appropriate value(s) along that line.

Tab. 1 Variables used for modelling. Explanations for terms; "Temp" = temperature, "stg1" = before aeration, "stg2" = after aeration, "stg3" = after secondary sedimentation, "TOC" = Total Organic Carbon, "N" = Nitrogen, "P" = Phosphorus, "SVI" = Sludge Volume Index, "DSVI" = Diluted Sludge Volume Index, "O2" = Oxygen concentration in aeration.

	Variable	Unit
1	Temp_raw_water	C°
2	Flow_stg3	m^3/d
3	Solids_stg1	g/m^3
4	pH_stg1	
5	Conductivity_stg1	mS/m
6	Temp_stg1	C°
7	COD_stg1	g/m^3
8	BOD_stg1	g/m^3
9	TOC_stg1	g/m^3
10	N_stg1	g/m^3
11	P_stg1	g/m^3
12	Sludgeload	kg BOD/kg MLSS/d
13	N:BOD	%
14	P:BOD	%
15	Sludge_age	d
16	Settling	ml/l
17	SVI	ml/g
18	DSVI	ml/g
19	Solids_stg2	g/m^3
20	O2_stg2	g/m^3
21	Ash_stg2	g/m^3
22	Sludge_blanket	%
23	Solids_stg3	g/m^3
24	pH_stg3	
25	COD_stg3	g/m^3
26	N_stg3	g/m^3
27	P_stg3	g/m^3
28	BOD_stg3	g/m^3
29	TOC_stg3	g/m^3

3.1.2 Filtering

Despite a large pre-sedimentation pool and an equalization basin, rapid changes in the quality of effluent may occur caused by an unexpected malfunction. However, the large volume of the sludge treatment process buffers the activated sludge and thus the quality of the activated sludge changes slowly. In addition, there is always some noise in accuracy of the data caused by sampling and laboratory measurements. For these reasons, the variables are filtered by a moving average filter, where the window size was ten days.

3.2 Self-Organizing Maps

Self-Organizing Maps (SOM) are an artificial neural network methodology, which can transform an n-dimensional input vector into a one- or two-dimensional discrete map. The input vectors, which have common features, are projected to the same area of the map e.g. (in this case described as "neurons"). Each neuron is associated with an n-dimensional reference vector, which provides a link between the output and input spaces. This lattice type of an array of neurons, which is called the map, can be illustrated as a rectangular, hexagonal, or even irregular organization. Nevertheless, the hexagonal organization is used most often, as it best present the connections between the neighboring neurons. The size of the map, as defined by the number of neurons, can be varied depending on the application; the more neurons, the more details appear.

At first, random values for the initial reference vectors are sampled from an even distribution, whereby the limits are determined by the input data. During learning, the input data vector is mapped onto a particular neuron based on the minimal n-dimensional distance between the input vector and the reference vectors of the neurons (Best Matching Unit, BMU). Then the reference vectors of the activated neurons are updated. When the trained map is applied, the best matching neurons are calculated using these reference vectors. In this unsupervised methodology, the SOM can be constructed without previous a priori knowledge [2].

The data were coded into 29 inputs for the SOM. All input values were variances scaled. The SOM having 256 neurons in a 16x16 hexagonal arrangement was constructed. The linear initialization and batch training algorithms were used in training the map. A Gaussian function was used as the neighborhood function. The map was taught with 50 epochs and the initial neighborhood had the value of 6. The SOM Toolbox program (v. 2.0 beta) was used in the analysis under a Matlab-software platform (Mathworks, Natick, MA, USA).

3.3 K-means

The K-means method is a well-known non-hierarchical cluster algorithm [10]. The K-means algorithm was applied to the clustering prototype vectors of the map. The basic version begins by randomly picking K cluster centers, assigning each point to the cluster whose mean is closest in a Euclidean-distance, then computing the mean vectors of the points assigned to each cluster, and using these as new centers in an iterative approach.

The number of clusters in the case specific application may not be known a priori. In the K-means algorithm the number of clusters has to be predefined. It is common that the algorithm is applied with different number of clusters and then the best solution among them is selected using a validity index [12]. The Davies-Bouldin (DB) index [11] is calculated as follow:

$$DB = \frac{1}{N}\sum_{i=1}^{N} \max_{jj\neq i} \frac{S_i + S_j}{d_{ij}} \quad (1)$$

where N is the number of clusters. The within (S_i) and between (d_{ij}) cluster distances are calculated using the cluster centroids as follows:

$$S_i = \frac{1}{|C_i|}\sum_{x\in C_i} \left\| x - m_j \right\| \quad (2)$$

$$d_{ij} = \left\| m_i - m_j \right\| \quad (3)$$

where m_i is the centre of cluster C_i, with $|C_i|$ the number of points belonging to cluster C_i. The objective is to find the set of clusters that minimizes Eq. (3).

The K-means algorithm was applied to the clustering of the trained map, or more precisely, to the clustering of the reference vectors. By clustering the map the interactions can be detected more easily, and the clusters can then be treated as sub-models of the main model, which was formed by the SOM-algorithm. The number of clusters was determined using the DB index. Smaller values of DB index indicate better clusters. After training and clustering, the desired reference vector elements of clustered neurons were visualized in a two-dimensional space to reveal the possible interactions between data variables.

4 Results and discussion

The map was obtained by training a self-organizing network using the data of the sludge treatment process as inputs. Component planes of the SOM model are shown in Fig. 2. The SOM was then clustered according to the reference vectors by using the K-means algorithm. The smallest value of DB index indicates the mathematical best number of clusters. In this case six clusters seemed to be competent as can be seen in Fig. 3. The map and the clusters with short descriptions are illustrated in Fig. 4.

Some interesting multidimensional correlations between certain process variables were found after clustering the modeled sludge treatment data. An example of these correlations is illustrated here. In Fig. 5, the neurons of the trained SOM map are presented according to the selected variable components of their reference vectors. The sludge settling is presented as a function of the sludge load in

Fig. 5. A tremendous variation between the behaviors of different clusters can be clearly observed. In the case of cluster 4, the sludge settling decreases with the growth of the sludge load. In contrast, in the case of cluster 2 the sludge settling increases with the sludge load. One reason for this different behavior can be seen in Fig. 6, where the sludge settling of both clusters is presented as a function of the temperature of the effluent. The temperature of the effluent and also the sludge age are clearly higger in the case of cluster 4 than in the case of cluster 2. These results indicate that this kind of approach is a useful way in modelling the sludge treatment process.

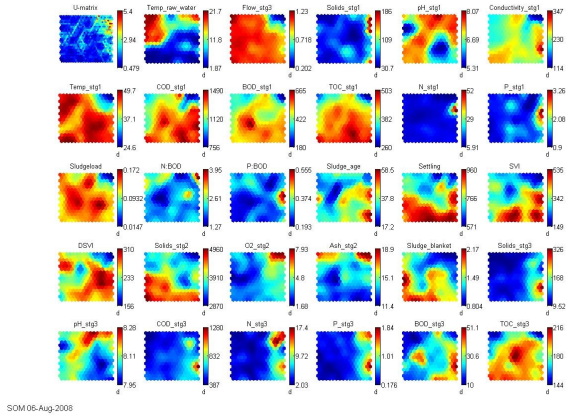

Fig. 2 Component planes of the SOM model.

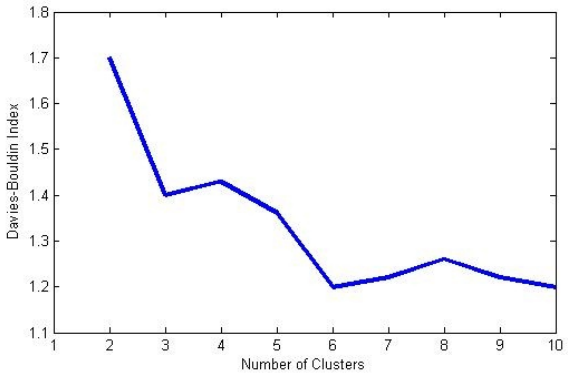

Fig. 3 Smaller values of Davies-Bouldin index indicate better clusters. In this case six clusters seemed to be competent.

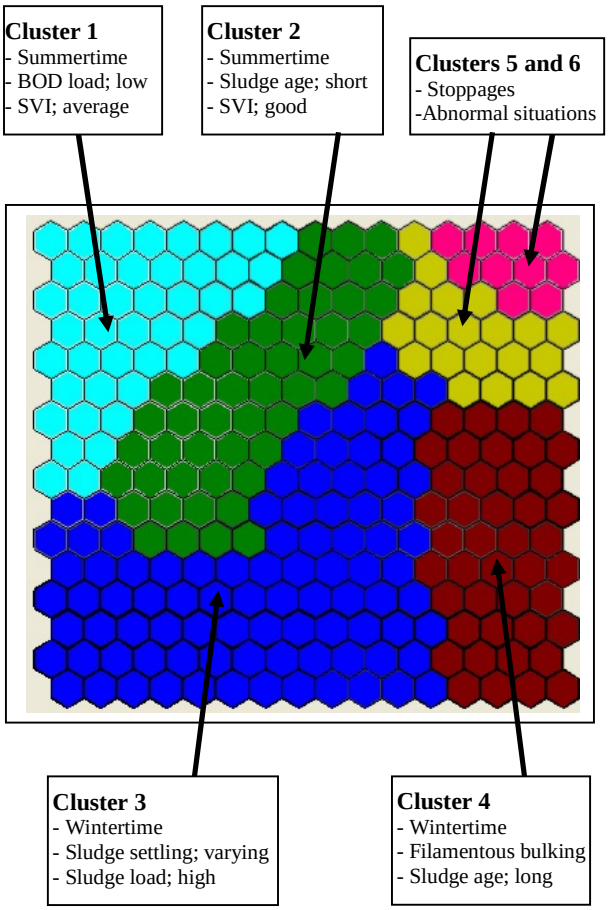

Fig. 4 Reference vectors of the SOM is clustered for six clusters by K-means method. Short descriptions for each cluster are also shown.

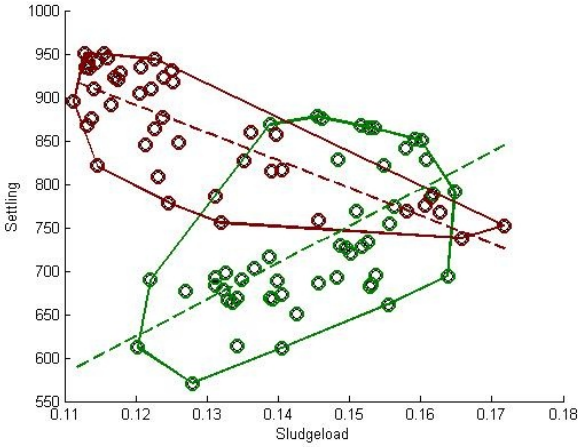

Fig. 5 Sludge settling presented as a function of the sludge load by using the reference vectors of neurons. The correlation is positive in Cluster 2 (green), but in Cluster 4 (brown) the correlation is negative instead.

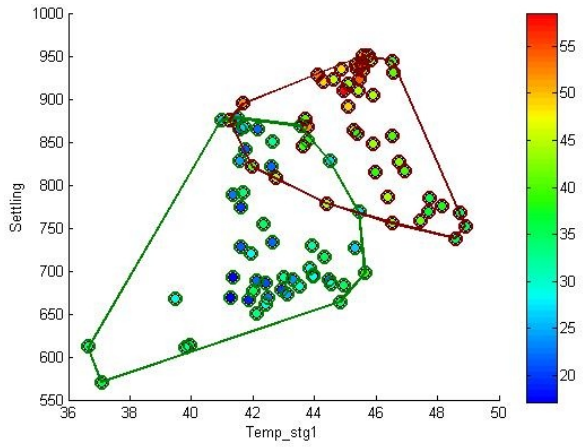

Fig. 6 Sludge settling of Cluster 2 (green) and Cluster 4 (brown) in a function of temperature of the effluent. The color bar represents sludge age.

The aim of the study was to discover whether the neural network modelling method could be a useful and time-saving way to analyse this kind of industrial process. The findings indicate that the approach described in this paper is a useful way to model the process data that were under examination. By using the selected method, interesting relations were found quite fast and easily, the relations which could have been much more difficult to find using traditional data processing methods

5 Conclusion

Because of the growing need for optimizing industrial processes due to, for example, environmental regulations of process, developing new methods for process analysis is very important. The results presented in this paper show that the applied SOM-based neural network method is an efficient and fruitful way to model data acquired from the sludge treatment. By means of this data-driven modelling method, some new findings were discovered concerning the dependencies between the process parameters.

6 References

[1] T. Becker, T. Enders and A. Delgado A. Dynamic neural networks as a tool for the online optimization of industrial fermentation, Bioprocess Biosyst Eng 24, Springer-Verlag Heidelberg, Germany, 347-354. 2002.

[2] Kohonen, T., Self-Organizing Maps, Springer-Verlag, Berlin Heidelberg, Germany, 2001.

[3] M. Heikkinen, A. Kettunen, E. Niemitalo, R. Kuivalainen and Y. Hiltunen. SOM-based method for process state monitoring and optimization in fluidized bed energy plant, Lecture Notes in Computer Science 3696, Eds.: W. Duch, J. Kacprzyk, E. Oja and S. Zadrozny, Springer-Verlag Berlin Heidelberg, Germany, 409-414. 2005.

[4] M. Heikkinen, V. Nurminen, T. Hiltunen and Y. Hiltunen. A Modeling and Optimization Tool for the Expandable Polystyrene Batch Process. Chemical Product and Process Modeling, 3(1), Article 3. The Berkeley Electronic Press. 2008.

[5] M. Heikkinen, T. Latvala, E. Juuso and Y. Hiltunen. SOM-based Modelling for an Activated Sludge Treatment Process. *The Institute of Electrical and Electronics Engineers IEEE*, Tenth International Conference on Computer Modelling and Simulation, EUROSIM/UKSim, Cambridge, UK, April 1-3 2008.

[6] M. Heikkinen, T. Heikkinen and Y. Hiltunen. Modelling of Activated Sludge Treatment Process in a Pulp Mill Using Neural Networks. The 6th International Conference on Computing, Communications and Control Technologies: CCCT 2008, Orlando, Florida, USA, June 29th – July 2nd 2008.

[7] M. A. Hussain. Review of the applications of neural networks in chemical process control: simulation and online implementation, Artificial intelligence in engineering, vol. 13, 1, Elsevier, Oxford, UK, 55-68. 1999.

[8] I. M. Mujtaba and M. A. Hussain. Application of Neural Network and Other Learning Technologies in Process Engineering, Imperial College Press, London, UK. 2001.

[9] C. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.

[10] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Statistics Vol. I, Berkeley and Los Angeles: University of California Press, 281–297, 1967.

[11] T. Räsänen, J. Ruuskanen and M. Kolehmainen. Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. *Applied Energy*, Vol. 85, 9, Elsevier, 830-840, 2008.

[12] D. Davies, D. Bouldin. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*. 2, 224–227, 1979.

7 Acknowledgements

This research was supported by the Finnish Funding Agency for Technology and Innovation (Tekes), Kemira Oyj, Stora Enso Oyj and UPM-Kymmene Oyj.