SOM-BASED SUBTRACTION ANALYSIS TO PROCESS DATA OF AN ACTIVATED SLUDGE TREATMENT PLANT

M. Heikkinen[1], H. Poutiainen[1], M. Liukkonen[1], T. Heikkinen[2], Y. Hiltunen[1]
[1]University of Kuopio, Kuopio, Finland; [2]UPM-Kymmene, Wisaforest, Pietarsaari, Finland

Corresponding Author: M. Heikkinen, University of Kuopio, Department of Environmental Science
P.O. Box 1627, FI-70211 Kuopio, Finland; mikko.heikkinen@uku.fi

Abstract. This paper presents an overview of an analysis method based on Self-Organizing Maps (SOM) which was applied to an activated sludge treatment process of the paper and pulp mill. 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 and to investigate process states. The used analysis procedure went as follows. At first, the process data is modelled using the SOM algorithm. Next, the reference vectors of the map were classified by K-means algorithm into four clusters, which represented different states of the process. At the final stage, the center vectors of the clusters were used for subtraction analysis to indicate differences between different process states. The results show that the method presented is an efficient way to analyze activated sludge process states and it could lead to better process control.

1 Introduction

Today, environmental regulations set challenges for controlling industrial emissions. Although the emission levels are nowadays low compared to the 1970´s for example, the continuing increase in industrial production makes it more and more important to use the best available technologies so that emissions can be minimized. In the area of industrial wastewater treatment the need to reduce emissions is further emphasized by the European Union directive 91/271 (Urban Wastewater Treatment-directive) that requires reductions in nitrogen loads from waste water to recipient water bodies.

Archived process data can be used 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 industrial processes [1, 2]. In this study we have constructed a self-organising map to optimizing an active sludge wastewater treatment process of an industrial pulp and paper mill.

2 Background and Methodology

Process. The UPM-Kymmene Wisaforest production site at Pietarsaari consists of a pulp mill, paper factory and a sawmill. These units use about 4.5 million m^3 of wood annually, and employ 650 persons. The annual production capacities of the Wisaforest pulp mill, paper factory and sawmill are 800 000 tons, 195 000 tons and 200 000 m^3, respectively. The wastewater streams of all units are handled in a single activated sludge wastewater treatment plant (WWTP). This unit treats an average of 86 000 m^3/day of waste water in a typical activated sludge process consisting of: (1) pre-sedimentation and equilibration, (2) aeration and (3) secondary clarifying (Figure 1).

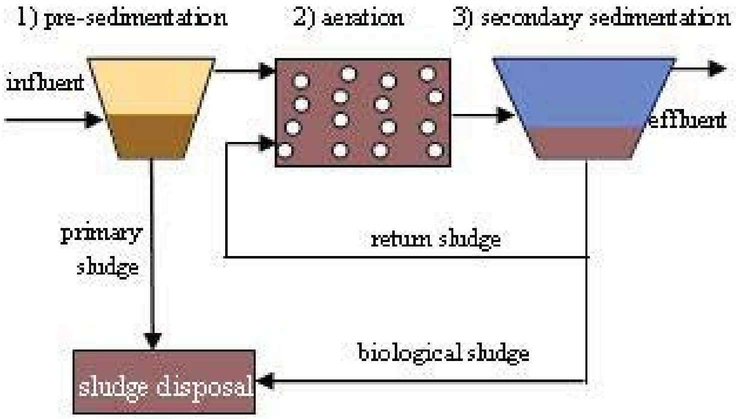

Figure 1. A simplified diagram of the activated sludge treatment process with three main stages; 1) pre-sedimentation, 2) aeration and 3) secondary sedimentation.

Data. The raw data (on-line data and laboratory measurements) was extracted from the databases of the pulp mill. The selection of variables that were used in the analysis was made by a process expert. The complete data set contained values of 29 variables for 4 years with one day resolution. The variables that were used in the modelling are presented in Table 1.

	Variable	Unit
1	Temperature, raw water	C°
2	Flow	m^3/d
3	Solids, influent	g/m^3
4	pH, influent	
5	Conductivity, influent	mS/m
6	Temperature, influent	C°
7	COD (Chemical Oxygen Demand), influent	g/m^3
8	BOD (Biological Oxygen Demand), influent	g/m^3
9	TOC, influent	g/m^3
10	N (nitrogen), influent	g/m^3
11	P (phosphorus), influent	g/m^3
12	Sludgeload, influent	kg BOD/kg MLSS/d
13	N:BOD, influent	%
14	P:BOD, influent	%
15	Sludge age	d
16	Sludge Settling	ml/l
17	SVI (Sludge Volume Index)	ml/g
18	DSVI (Diluted Sludge Volume Index)	ml/g
19	Solids, after aeration	g/m^3
20	Oxygen, in aeration	g/m^3
21	Ash, after aeration	g/m^3
22	Sludge blanket, secondary clarifier	%
23	Solids, effluent	g/m^3
24	pH, effluent	
25	COD, effluent	g/m^3
26	N, effluent	g/m^3
27	P, effluent	g/m^3
28	BOD, effluent	g/m^3
29	TOC (Total Organic Carbon), effluent	g/m^3

Table 1. Variables used for modelling.

Pre-processing the data. Two types of missing data were observed in the on-line and laboratory measurement sets. In the on-line measurement set the missing data had been caused by measurement errors and measurement equipment malfunctions. In the laboratory measurement set the missing data had 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 neighbouring values and returning the appropriate value(s) along that line.

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

Self-Organizing Map. Self-Organizing Map (SOM) is a well-known unsupervised learning algorithm [3]. SOM 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 presents the connections between the neighbouring 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 [3].

In this study, the SOM had 144 neurons in a 12x12 hexagonal arrangement. All input values were variances scaled. The linear initialization and batch training algorithms were used in training the map. A Gaussian function was used as the neighbourhood function. The map was taught with 100 epochs and the initial neighbourhood 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).

Clustering method. In this paper, the K-means algorithm was applied to the clustering reference vectors of the map. The K-means method is a well-known non-hierarchical cluster algorithm [4]. The basic version begins by randomly picking K cluster center, assigning each point to the cluster whose mean is closest in a Euclidean-distance, and then computing the mean vectors of the points assigned to each cluster, and using these as new center in an iterative approach.

Subtraction analysis of reference vectors. Each reference vector, which represents the common features of the data in each neuron, is defined during the training of the map. In the subtraction analysis, the reference vectors of two neurons are subtracted from each other. This method can be used for identification of any differences in factors between corresponding subgroups of two neurons or clusters. In the case of identifying differences of clusters, the center vector of the clusters which are solved by K-means algorithm, can be used for the subtraction.

3 Results

The SOM was obtained by training a self-organizing network with the data of an activated sludge treatment process. Component planes of the SOM model are shown in Figure 2. Four clusters, calculated by the K-means method, are shown in the clustered map (Figure 3). The center vectors of each cluster, illustrated by the bar graphs in Figure 3, show the features of the clusters.

Subtraction analysis of the vectors of the center of clusters revealed differences in process factors between clusters. In Figure 4 is an illustrated subtraction between clusters 1 and 3.

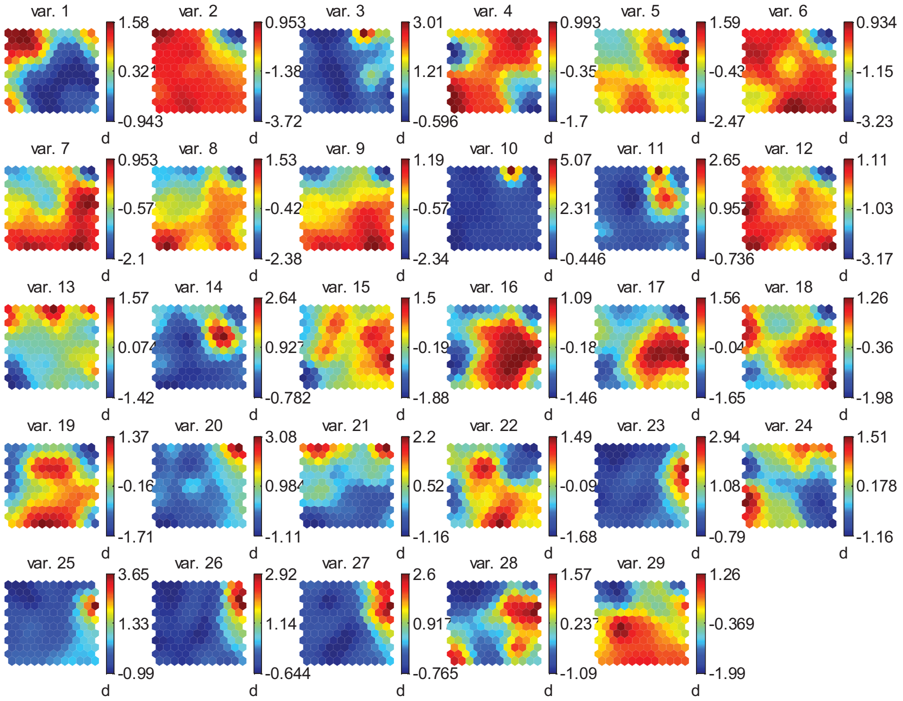

Figure 2. Component planes of the SOM model. Values of the reference vectors are variance scaled.

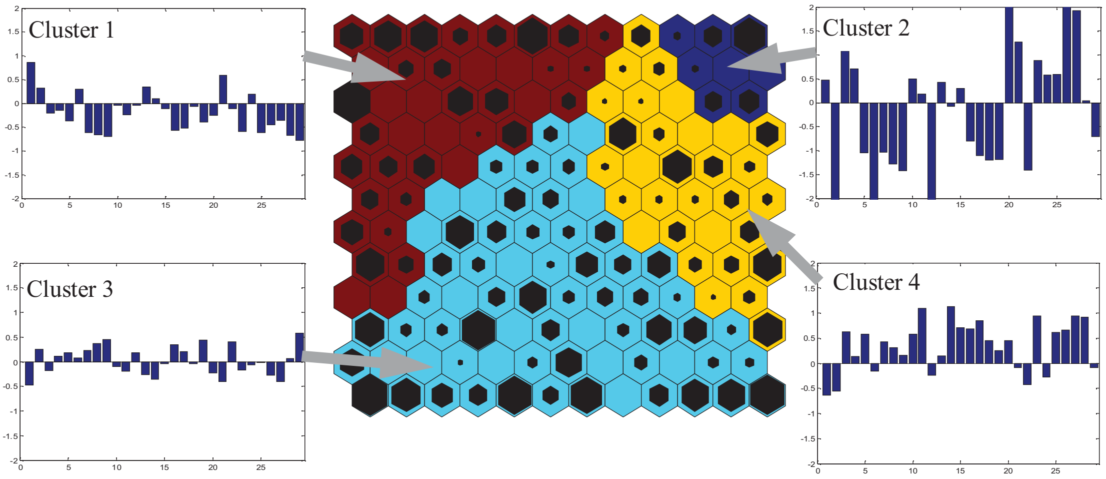

Figure 3. SOM using the data of activated sludge treatment process showing the number of the hits on the size of the depicted neuron. The background colours show the four main clusters of the map. The bar graphs represent the reference vectors related to four clusters.

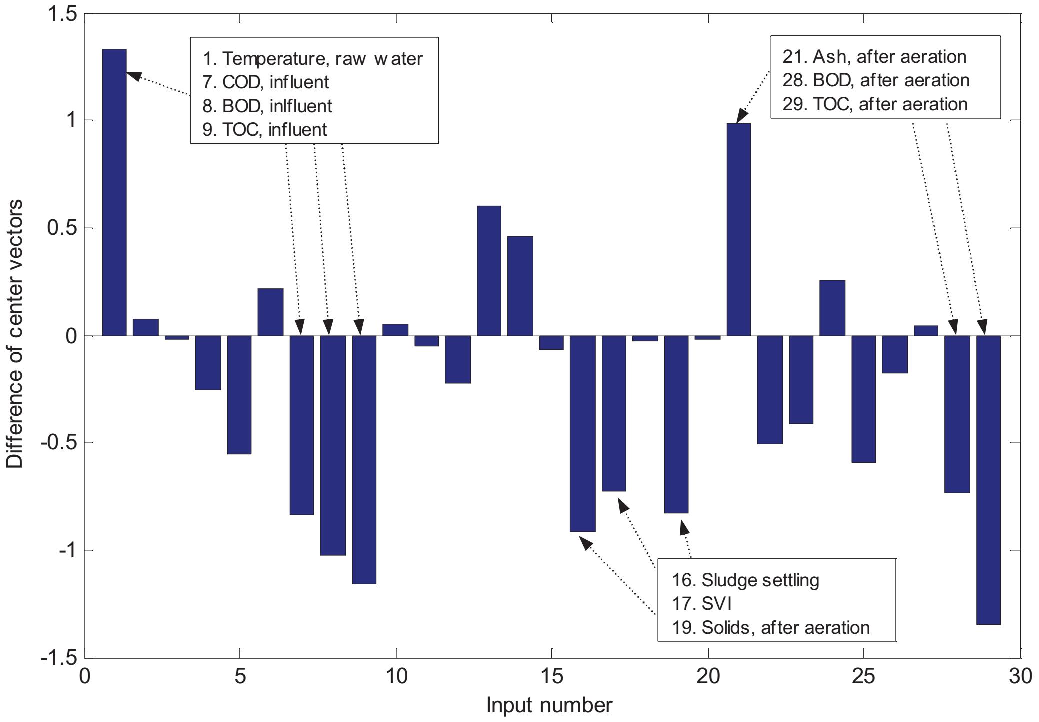

Figure 4. The bar graph represents the difference between center vector of cluster 1 and the center vector of cluster 3. The reference vectors of the clusters are subtracted from each other (the values are variance scaled). The most remarkable differences are shown in the figure.

4 Discussion

The aim of the study was to apply Self-Organizing Maps (SOM) for optimizing the active sludge wastewater treatment process of an industrial pulp and paper mill. The SOM described in this study was trained using a four-year data set of the activated sludge process. The SOM was clustered according to the reference vectors by using the K-means algorithm. The map and the clusters with short descriptions are illustrated in Figure 5.

- Cluster 1 is prevailing in the summertime when the organic load of influent is lower than average. In addition, in this cluster the sludge settling properties are good and emissions to receiving waterways are low.

- Cluster 2 represents periods when the whole process has either been shut down, or it has been out of control (abnormal situations). This process state will not be considered from the point of view of optimization.
- Cluster 3 is common in the wintertime, when the organic load of influent is high. The sludge settling velocity is varying and the concentration of solids in aeration is high.
- Cluster 4 represents a process state, where the process is somewhat unstable. Filamentous sludge is often encountered, and the problems associated with it like bulking and foaming may cause higher emission levels than normal in the effluent. Cluster 4 is prevalent in wintertime, i.e. the process seems to be more difficult to control in the winter. Here also high sludge age is often encountered.

If we compare cluster 1 and cluster 3 by subtraction analysis (Figure 4), we see that the factors representing sludge settling properties, i.e. settling (variable 16) and SVI (Sludge Volume Index, variable 17) are better in cluster 3 than in cluster 1. Also values of COD, BOD and TOC are lower in cluster 3 than in cluster 1. These clusters differ also in respect to temperature and time of year. High temperature may be the reason for poor sludge settleability, but also one of many influencing factors. For example it is known, that the composition of wood is different in wood harvested in winter compared to wood harvested in summer [5]. Both the equilibrium moisture content and (Brinell) hardness of wood differ. Wood felled in winter has also been found to have a higher density [6] and to contain more fatty- and resin acids than wood felled in summer [7, 8]. The effect of these varying wood properties and resulting seasonal variations in active sludge process influent from a pulp and paper mill has not been studied extensively.

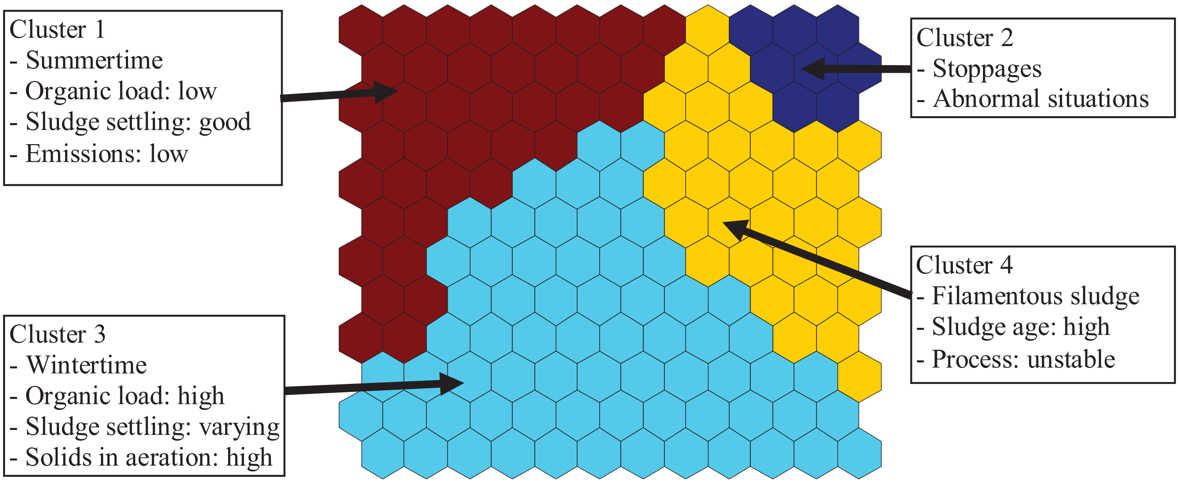

Figure 5. SOM is clustered for four clusters. Short descriptions for each cluster are also shown.

5 Conclusions

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 activated sludge treatment process. By means of this data-driven modelling method, some new findings were discovered concerning the dependencies between the process parameters.

6 Acknowledgements

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

7 References

[1] Hussain, M. A.: *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, (1999), 55 - 68.

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

[3] Kohonen, T.: *Self-Organizing Maps*. Springer-Verlag, Berlin Heidelberg, Germany, (2001).

[4] MacQueen, J.: *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, (1967), 281 – 297.

[5] Möttönen, V., Heräjärvi, H., Koivunen H. and Lindblad, J.: *Influence of felling season, drying method and within-tree location on the Brinell hardness and equilibrium moisture content of wood from 27-35-year-old Betula pendula*. Scandinavian Journal of Forest Research, Vol. 19, 3, (2004), 241 – 249.

[6] Möttönen, V. and Luostarinen, K.: *Variation in density and schrinkage of birch (Betula pendula Roth) timber from plantations and naturally regenerated forests*. Forest Products Journal 2006, Vol. 56, 1, (2006), 34 - 39.

[7] Mirza, S., Harvey, G., Sénéchal, M. and Quellet, S.: *The use of lipase enzymes in TMP furnish: A practical perspective*. TAPPSA Journal, The Technical Association of The Pulp and Paper Industry of South Africa, (presented at Paptac, September 2006).

[8] Nerg, A., Kainulainen, P., Vuorinen, M., Hanso, M., Holopainen, J.K. and Kurkela, T.; *Seasonal and geographical variation of terpenes, resin acids and total phenolics in nursery grown seedlings of Scots pine*. (Pinus sylvestris L.), New Phytol. 128, (1994), 703 - 713.