Self-Organizing Map based operating regime estimation for state based control of Wastewater Treatment Plants

Peter Kern,
Christian Wolf, Michael Bongards
Institute for Automation & Industrial IT
Cologne University of Applied Sciences
Cologne, Germany
peter.kern@fh-koeln.de, christian.wolf@fh-koeln.de
michael.bongards@fh-koeln.de

Tosin Daniel Oyetoyan
Software Engineering Group
Department of Computer and Information Science (IDI)
Norwegian University of Science and Technology
Trondheim, Norway
tosindo@idi.ntnu.no

Seán McLoone
Department of Electronic Engineering
National University of Ireland Maynooth
Maynooth, Ireland
sean.mcloone@eeng.nuim.ie

Abstract— **An optimal control of wastewater treatment plants (WWTP) has to account for changes in the bio-chemical state of the bioreactors. As many process variables of a WWTP are not measurable online, the development of an efficient control strategy is one of the greatest challenges in the optimization of WWTP operation. This paper presents an approach, which combines the use of Self-Organizing Maps (SOM) and a clustering algorithm to identify operational patterns in WWTP process data. These patterns provide a basis for the optimization of controller set points that are well suited for the previously identified operation regimes of the plant. The optimization is performed using Genetic Algorithms. This approach was developed, tested and validated on a simulation model based on the Activated Sludge Model No.1 (ASM1). The results of this state-based control indicate that the presented methodology is a promising and useful control strategy that is definitely able to address the distinctive energy and effluent limit challenges faced by WWTP operators.**

Wastewater Treatment; State based Control; Self Organizing Maps; Clustering; Optimization, Genetic Algorithm

I. INTRODUCTION

Rising energy prices as well as stricter governmental regulations for the effluent quality of wastewater treatment plants require sophisticated control strategies to operate plants as efficiently as possible. In general, wastewater treatment consists of several bio-chemical processes, which take place sequentially and simultaneously inside the bioreactors of a WWTP. These processes are nitrification for ammonium (S_{NH}) removal, denitrification for nitrate (S_{NO}) removal, carbon degradation, hydrolysis, etc. Out of these processes, most of the energy consumed in WWTPs is needed for the nitrification process, which takes place in permanently aerated bioreactors. This aeration is usually performed by huge compressors, resulting in high energy consumption. For this reason the intelligent control of the oxygen concentration (S_O) inside the aerated bioreactors can save up to 20% energy in comparison with standard basic controllers [1].

This paper presents an approach to finding the optimal oxygen concentration for ammonium removal as efficiently as possible depending on the current state of the plant. The computer experiments were conducted with a simulation model developed with the commercial Matlab toolbox Simba [2]. The model represents a simple WWTP with two bioreactors as described in section III. Using a SOM on the generated training data from measurable plant variables, a trained map is obtained. Further clustering of this feature map, provides a final map where each feature represents an actual state operating regime of the WWTP. Thus, the current operating regime of the plant can be easily predicted using the final feature map. Subsequently, for each operating regime, an optimal set point for the S_O controller is determined by exploring the search space with Genetic Algorithms.

II. CHALLENGES IN WWTP CONTROL

Most WWTP are operated with simple controllers using fixed oxygen set points [1]. More sophisticated controllers calculate the necessary oxygen concentration based on at least one process parameter, usually the ammonium concentration. This results in better control performance but is still a vast approximation to the real challenge. To determine the optimal oxygen concentration for the bioreactors in order to get the highest ammonium degradation ($D_{S_{NH}}$), the state of the bioreactor needs to be fully-known. In practice, this is not possible.

During the nitrification process, ammonium (S_{NH}) is oxidized to nitrate (S_{NO}), so that oxygen (S_O) is consumed. Under the theoretical assumption that all state variables except ammonium concentration (S_{NH}) and oxygen concentration (S_O) are constant, it is possible to plot the

978-1-4577-1196-1/11/$26.00 ©2011 IEEE

general shape of the ammonium degradation function for a certain state. In reality, shape and position of the best area depend on several other state variables. Figure 1 shows the area, where the ammonium degradation is relatively high, while the oxygen concentration is still adequate (<2.5mg/l). It is evident, that there is always a trade-off between oxygen concentration, meaning higher energy consumption, and ammonium degradation. Therefore, the most efficient ratio between oxygen concentration (S_O) and ammonium (S_{NH}) to get the most efficient ammonium degradation ($D_{S_{NH}}$) with the least amount of energy, can be determined.

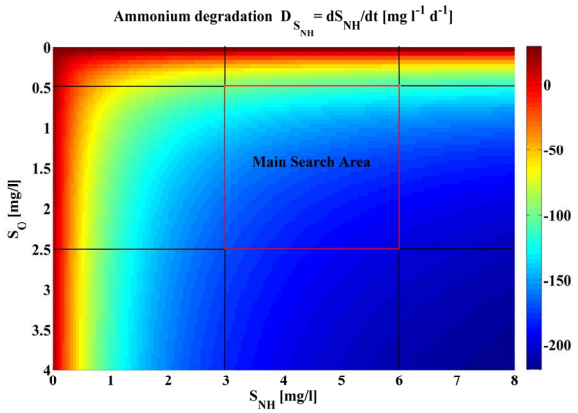

Figure 1. Ammonium degradation

III. Description of the Simulation Model

The simulation model used is based on the Activated Sludge Model 1 (ASM1) developed by the International Water Association (IWA). ASM1 describes the bio-chemical processes inside the bioreactors of a WWTP using 8 coupled linear differential equations. The internal state of each bioreactor is described by 13 concentration components and the hydraulic flow [3].

- S_I - soluble inert organic matter [mg/l]
- S_S - readily biodegradable substrate [mg/l]
- X_I - particulate inert organic matter [mg/l]
- X_S - slowly biodegradable matter [mg/l]
- $X_{B,H}$ - active heterotrophic biomass [mg/l]
- $X_{B,A}$ - active autotrophic biomass [mg/l]
- X_P - particulate products arising from biomass [mg/l]
- S_O - oxygen [mg/l]
- S_{NO} - nitrate and nitrite nitrogen [mg/l]
- S_{NH} - ammonium and ammonia nitrogen [mg/l]
- S_{ND} - soluble biodegradable organic nitrogen [mg/l]
- X_{ND} - particulate biodegradable nitrogen [mg/l]
- S_{ALK} - alkalinity [mole $HCO_3^-m^3$]
- Q - hydraulic flow [m^3/d]

The simulation model used for this simulation study represents a WWTP with two biological reactors (see Figure 2). An upstream denitrification tank with a size of 2000m^3 is followed by a nitrification tank with a size of 4000m^3 and a clarifier with a surface area of 1000m^2. The nitrate concentration is controlled by internal recirculation. The model is based on the Benchmark Simulation Model No.1 (BSM) [4] with five bioreactors, which represents a typical European WWTP.

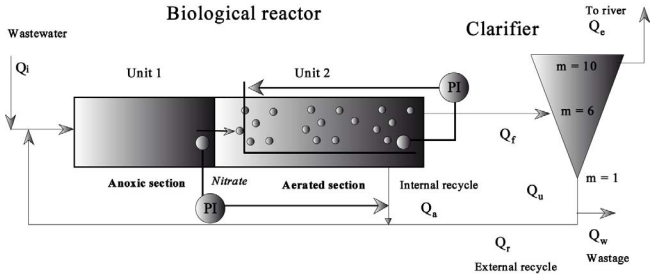

Figure 2. Simulation Model

The standard aeration control in the simulation model operates with a fixed set point of 2mg/l for the oxygen concentration. This value is considered a good compromise between energy consumption and effluent quality for this plant under the given inflow conditions. The IWA Taskgroup on Benchmarking of Control Strategies developed several typical inflow scenarios for the BSM 1. Figure 3 depicts the dry weather inflow scenario for hydraulic flow and S_{NH}.

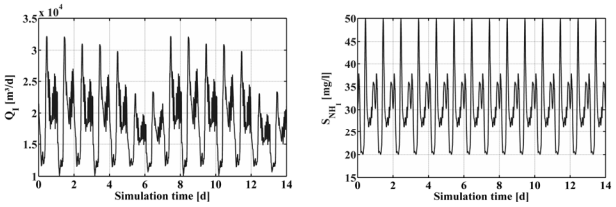

Figure 3. Dry weather inflow scenario

Due to the fact that the size of the plant is similar to the BSM1, these scenarios are also realistic for the model and therefore are used as inflow.

A. Model Initialization

To initialize the model, a 100-days period of stabilization in closed-loop using constant inputs (average dry weather inflow) is simulated. The system stabilizes during this period. This period is particularly important for the growth of the biomass $X_{B,H}$ and $X_{B,A}$ and for achieving a reference steady state of the plant for further simulations. Following this, a 14-day simulation period with dry weather inflow is performed to bring the plant to a desired state for testing the controller strategy.

IV. Data Generation and Variable Selection

The data for the SOM clustering is created synthetically using the initialized and fully calibrated simulation model as

described in section two. To make the SOM clustering as generally applicable as possible, three different weather case scenarios were simulated for 14 days each, namely dry, rainy and stormy weather. During these simulation runs, the state vector of the ASM1 for the aerated bioreactor is sampled every 15 minutes for each weather scenario. Due to the fact, that not all 14 process variables of the state vector are relevant for the aeration control of the bioreactor, the most suitable variables have to be selected. Furthermore, it is important to consider the fact that some process variables are extremely difficult to measure, which reduces practical applicability of the presented optimization approach. Therefore, selection is performed based on measurability and relevance of the variables for the nitrification process in the aerated bioreactor as stated in [5]. The selected variables are S_{NH}, S_{NO}, Q and the sum of S_S and X_S, which represents the portion of available degradable Chemical Oxygen Demand (*COD*). In addition to these state vector variables, an approximation of the oxygen consumption O_C in the bioreactor is used. This is defined as the total airflow into the bioreactor Q_{air} divided by the oxygen concentration S_O inside the bioreactor.

$$O_C = \frac{Q_{air}}{S_O} \tag{1}$$

A. Data Preprocessing

As a preprocessing step, the data is scaled between 0 and 1 using the min-max method. During the training of the SOM, Euclidean distance is used to determine the similarity or dissimilarity between the nodes and the input vector. Without normalization, high values from variables like *COD* or *Q* would have significantly higher influence on the SOM.

V. Self Organizing Maps

Self-Organizing Map (SOM) is a special kind of an artificial neural network, whose training is unsupervised and which has properties of vector quantization and vector projection algorithms. SOM reduces multidimensional data into much lower dimensional spaces, usually two dimensions. This method of dimension reduction, called vector quantization, tries to find a prototypical representation of the original data [6]. In addition, it performs a vector projection, which creates a topology preserving mapping from the high dimensional input space to the two dimensional output space usually referred to as the map. This in effect means that data that are close together in the input space are mapped into a spatially close area on the map and elements which are spatially close together on the map should be similar in the input space [7].

The SOM algorithm can be divided into three parts; the architecture, initialization and the training. The SOM consists of artificial neurons randomly created and fitted to the lattice of a map. Each neuron contains a vector of weights $\boldsymbol{m}_i = [m_{i1},...,m_{in}]$, i=1,…,Q where n is the dimension of the weight/input vector and Q is the number of the map nodes. Each input vector $\boldsymbol{x}_j = [x_{j1},...,x_{jn}]$, j=1,…, P where P is the size of the input dataset is mapped during training to exactly one neuron $\boldsymbol{m}_i$. The initialization is done by randomly initializing the network with uniformly distributed values or by sample initialization with random samples drawn from the training set [6]. During training, each single neuron is activated and the best matching unit (BMU) to the input vector $\boldsymbol{x}_j$ is determined by a distance measure. For this work, Euclidean distance is used, it is computed as

$$d(\boldsymbol{x}_j, \boldsymbol{m}_i) = \sqrt{\sum_{k=1}^{n} (x_{jk} - m_{ik})^2}\,. \tag{2}$$

An adaptation rule for the neuron weight $\boldsymbol{m}_j$ is defined by:

$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t) + h_{ci} \cdot [\boldsymbol{x}_j(t) - \boldsymbol{m}_i(t)] \tag{3}$$

where h_{ci} is the neighborhood of BMU $\boldsymbol{m}_{ci}$ at time t and it defines the region of influence the input sample $\boldsymbol{x}$ has on the SOM.

Regarding the control application, SOM has several advantages compared to other methods. First, SOM is relatively insensitive to non-equally distributed data, which is important if certain operating regimes appear less frequently. Due to the fact that similar data will always hit the same BMU, all operating conditions are still represented by their BMUs and not masked as done by other clustering algorithms.

Furthermore, in terms of visualization, SOM clusters are graphically well represented. Due to the applied training method, the resulting data clusters are presented in a structured way showing the topology preservation feature.

A. Clustering the Map

For the purpose of this work, it is important to create bigger clusters from the SOM to further group similar operating regimes together. Different algorithms such as single linkage, complete linkage, ward linkage, centroid linkage and k-means can be applied on the SOM to create such bigger clusters. From experiments conducted, ward linkage fared better than other algorithms. Ward linkage aims to keep the variance between the clusters as low as possible. Mathematically, the distance in ward linkage clustering is computed as

$$d(C_1, C_2) = \frac{\left\| \boldsymbol{c}_1 - \boldsymbol{c}_2 \right\|^2}{\frac{1}{n_{C1}} + \frac{1}{n_{C2}}} \tag{4}$$

where C_1 and C_2 represent clusters 1 and 2, $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are the centroids of the clusters and n_{C1} and n_{C2} are the number of points in the clusters.

VI. SOM Analysis and Validation

In this section, the SOM is optimized based on two quality factors, the quantization error and the topographic error. The quantization error is defined as the sum of Euclidean distances of all input vectors to the weight vectors of the best matching unit (also known as the winning neurons) divided by the number of input samples. This is defined as

$$\varepsilon_Q = \frac{1}{P}\sum_{j=1}^{P} d(\boldsymbol{x}_j, \boldsymbol{m}_{cj}) \tag{5}$$

where $\boldsymbol{m}_{cj}$ is the BMU for the corresponding $\boldsymbol{x}_j$.

The projection quality is defined by the topographic error ε_T and is computed as follows. For all input data vectors, the nearest weight vector (BMU) and the second-nearest weight vector (Second BMU) are computed. If they are not adjacent on the SOM-grid, this is counted as a local error. For the global topographic error measure, the number of local errors is summed up and divided by the overall number of data samples [8].

$$\varepsilon_T = \frac{1}{P}\sum_{j=1}^{P} u(\boldsymbol{x}_j) \tag{6}$$

The topographic error is calculated as shown above where the function $u(x_j)$ is 0 if $\boldsymbol{x}_j$ data vector's first and second BMUs are adjacent and, 1 otherwise [9].

The best SOM for the given dataset is realized by using GA to determine the optimal map dimension, learning rate and number of iterations. The optimization approach is described in Figure 4. The Java SOMToolBox developed by the Technical University of Vienna [10] was extended and is used for this work. To integrate both platforms for the optimization steps, an interface between the Matlab GA toolbox (also the simulation model) and the Java SOMToolBox was developed.

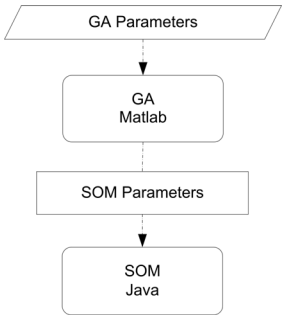

Figure 4. GA SOM Optimization

Given that the cost functions are quantization and topographic errors, objective functions and boundaries are defined as follows

$$\begin{aligned} Cost_a &:= \max(\varepsilon_T) \\ Cost_b &:= \min(\varepsilon_Q) \end{aligned} \tag{7}$$

subject to :

- $d_1 \le d \le d_2$, where d_1 and d_2 are the min & max map dimensions
- $lr_1 \le lr \le lr_2$, where lr_1 and lr_2 are the min & max map learning rates
- $itr_1 \le itr \le itr_2$, where itr_1 and itr_2 are the min & max map iterations.

The objective function is formulated as a single-objective function for each cost function separately, because the computational effort for a combined multi-objective function can grow significantly. Table I and II show the optimization parameters and the results respectively.

TABLE I. GA-SOM PARAMETERIZATION

Parameters			
Boundary	$d[x, y]$	lr	itr
Lower	20	0.5	90000
Upper	60	1.0	150000

TABLE II. GA-SOM OPTIMIZATION RESULTS

Optimization results				
SOM tuned parameters			***Quality factors***	
$d[x, y]$	lr	itr	ε_Q	ε_T
58,58	0.9561	147560	0.0343	0.1182
52,52	0.9785	131720	0.0377	0.1033

Table II shows that the results for ε_Q and ε_T differ by approximately 10%. The results for ε_T are used due to the fact that topology preservation is more important for this work as it ensures that similar operation regimes are close to each other.

VII. Prediction Model

The SOM is trained with the optimal parameters as described in section VI. The model generated is further clustered using a ward linkage algorithm in order to create the state prediction model. From several experiments conducted, ward linkage proved better in comparison to complete linkage and k-means. To validate these results, the silhouette algorithm [11] is applied on the data and predicted states. The mean values of the results are given in Table III.

TABLE III. MEAN VALUE OF SILHOUETTE ALGORITHM APPLIED TO SOM

SOM Prediction Model with 5 clusters			
Method	Ward	Complete	K-means
Mean	0.5011	0.3895	0.3677

Figure 5 shows the steps for generating the prediction model.

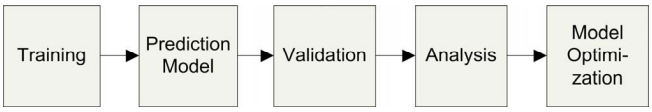

Figure 5. Prediction model design steps.

VIII. OPERATIONAL STATE-BASED CONTROLLER DESIGN

The SOM prediction model developed is able to separate different operation regimes of the plant. This adaptive ability can be used as part of a S_O controller. The model is fed continuously with the current measurable state variables and the current operation regime is predicted. Using a look up table each operation regime is assigned to its optimized S_O set point found by GA, as described in the subsequent section.

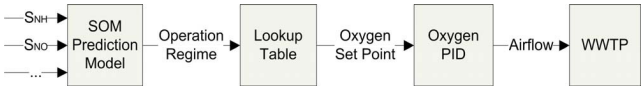

Figure 6. SOM Control Principle

A. Optimization of Set Points

The optimization aims to find an optimal S_O set point for each operation regime. To find this, a simulation experiment with a fixed S_O set point of 2mg/l for the inflow scenario was applied. By applying the SOM during the plant's continuous operation, the operation regime is predicted at each point in time. The produced data was reviewed for the longest continuous appearances of each state.

B. Fitness Function

For the optimization of the S_O set points genetic algorithms (GA) are used. The fitness function for the optimization is defined as:

$$f := \min\left(\int_{t_{S_i}}^{t_{E_i}} \frac{Q_{air}}{D_{S_{NH}}} dt\right) \quad (8)$$

Where Q_{air} is the air pumped into the bioreactor, $D_{S_{NH}}$ is the S_{NH} degradation, t_{S_i} the start time and t_{E_i} the end time of state i. This function has a minimum where the ratio between airflow and ammonium degradation is optimal. A well dimensioned WWTP is able to keep the desired effluent values at the point of highest energy efficiency. This means that the fitness function has to be modified for overloaded or under loaded WWTP.

IX. RESULTS AND DISCUSSION

In this section, the various results from several experiments conducted are discussed. Figure 7 shows three operational states identified by the SOM prediction model. The model is formulated based on the clustered SOM as shown in Figure 8. As discussed in section VII, the map shows the result of clusters produced by the Ward linkage algorithm and the various islands are generated with a Smoothed Data Histogram (SDH) [12]. The SDH identifies clusters by resembling the probability distribution of the data on the map. Due to the complex character of the plant's states, it is a challenge to relate the state to certain measurement values. Looking at Figure 7, it becomes obvious that the states follow a daily course. This daily course represents a typical load of a WWTP. From this, it can be argued that the determined state represents the state of the plant. A typical operational pattern of the plant revealed that in the night the plant operates at the minimal load, while the load is rated medium in the morning and highest around mid-day. The pattern formed by the clusters (see Figure 7) showed a correlation with the operational load. Cluster 1 represents the night load, cluster 2, reveals the transition period load and cluster 3 represents the mid-day load. From figure 8, it can be seen that the ward linkage algorithm separates the clusters at the same boundaries as the SDH algorithm. The three clusters correspond to the inflow categories of low, medium and high as described in [13].

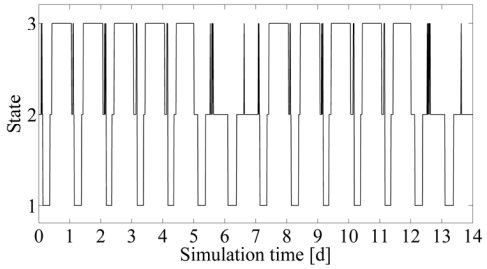

Figure 7. Recorded States

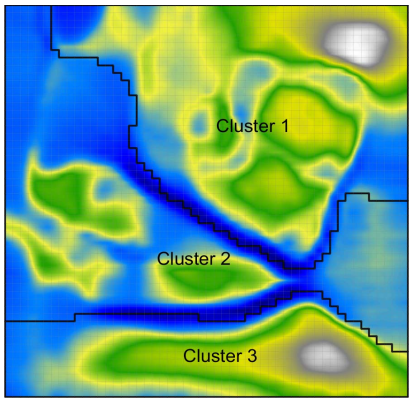

Figure 8. SOM clusters with SDH and Ward Linkage

In Table IV, the optimized O2 set points found by GA are tabulated. It is noticeable that for cluster 1 during the night a high set point was found. This is because the set point does not directly represent the airflow into the plant, rather the most efficient area.

TABLE IV. OPERATION REGIME SET POINTS

Optimization Results	
Operation Regime	***Optimized S_O Set Point [mg/l]***
1	2.48
2	1.77
3	2.26

Furthermore, the mean value for S_O of the three operation regimes is 2.17mg/l. This value is, as expected, very close to the fixed 2mg/l set point which is considered optimal for the BSM1 plant. Although the effluent values were not included in the fitness function, good results with regard to efficient plant operation and effluent values were achieved by the presented methodology. Keeping the effluent values and the energy in a reasonable domain shows that the plant is well operated. Otherwise the effluent values would be violated or the overall energy consumption would increase significantly. Table V gives an overview of the most important effluent parameters as well as the aeration energy. The aeration energy was calculated for the second week of the simulation period as described in the simulation procedure for the BSM1 [4]. The results show that the aeration energy and the total nitrogen are kept nearly at the same level while yielding a reduction of 3.3% in ammonium in the effluent.

TABLE V. PLANT PERFORMANCE

Optimization Results		
	Fixed S_O Set Point	***Optimized S_O Set Point***
Total Nitrogen	17.97	18.06 mg/l
Ammonium	3.56 mg/l	3.44 mg/l
Aeration Energy	1163 kWh	1184 kWh

X. CONCLUSION

The challenges of controlling a highly disturbed, non-linear system like a WWTP are not trivial. This work introduces new possibilities for control applications in wastewater treatment. Although, the favored results are very similar to the static S_O set point controller results, many potential opportunities are created to achieve improved results in this field. An important aspect of this work dealt with operation state identification from the process variables of a WWTP. The results proved encouraging with the clustering algorithm employed capturing and successfully categorizing the operating states of the WWTP. This in itself provides a platform for exploring its integration to developing control strategies. The outcome of this study presents a great potential for discovering some level of optimum in controlling the energy usage and at the same time keeping the effluent limits given by regulations. In conclusion, a framework for state-based data driven controller design has been developed. This strategy is promising and offers great potential for achieving optimum energy control and at the same time compliance with effluent limits.

XI. REFERENCES

[1] G. Olsson, M. K. Nielsen, Zhiguo Yuan, A. Lynggaard-Jensen, and J.-P. Steyer, Instrumentation, control and automation in wastewater systems. London: IWA Pub, 2005.

[2] SIMBA-Simulation of the biological wastewater treatment, ifak Magdeburg e.V., Availableat http://www.ifak.eu (2011, Jun. 13).

[3] M. Henze, Activated sludge models ASM1, ASM2, ASM2d and ASM3. London: IWA Publ, 2007.

[4] J. Alex, L. Benedetti, J. Copp, K. Gernaey, U. Jeppson, I. Nopens, M.-N. Pons, L. Rieger, C. Rosen, J. Steyer, P. Vanrolleghem, and S. Winkler, "Benchmark Simulation Model no. 1 (BSM1)," Lund University, Lund, 2008.

[5] M. Henze, Biological wastewater treatment: Principles, modelling and design. London: IWA Publ, 2008.

[6] A. P. Engelbrecht, Computational intelligence: An introduction, 2nd ed. Chichester: Wiley, 2007.

[7] T. Kohonen, Self-organizing maps: With 22 tables, 3rd ed. Berlin, Heidelberg: Springer, 2001.

[8] D. Baum, "Visualization for Comparing Self-organizing Maps," Master Thesis, Vienna University of Technology, 2007.

[9] E. Arsuaga Uriarte and F. Diaz Martin, "Topology Preservation in SOM," International Journal of Mathematical and Computer Sciences, no. 1:1, pp. 19–22, 2005.

[10] The Java SOMToolbox @ IFS, TU Vienna. Available: http://www.ifs.tuwien.ac.at/dm/somtoolbox/index.html (2010, Jun. 28).

[11] L. Kaufman, Finding groups in data: An introduction to cluster analysis. Hoboken, N.J: Wiley, 2005.

[12] E. Pampalk, A. Rauber, and D. Merkl, "Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Maps," in In Proceedings of the International Conference on Artifical Neural Networks ICANN'02: Springer, 2002, pp. 871–876.

[13] A. Ebel, "Application of Computational Intelligence Techniques to Modelling, Control and Optimisation of Wastewater Treatment Plants," Doctoral Dissertation, Department of Electronic Engineering, National University of Ireland, Maynooth, Maynooth, 2009