Monitoring a Secondary Settler using Gaussian Mixture Models

Jesús Zambrano[1] Oscar Samuelsson[2,3] Bengt Carlsson[3]

[1]School of Business, Society and Engineering, Mälardalen University, Box 883, 72123 Västerås, Sweden,
jesus.zambrano@mdh.se
[2]IVL Swedish Environmental Research Institute, P.O. Box 210 60, 10031 Stockholm, Sweden.
[3]Department of Information Technology, Uppsala University, Box 337, 75105 Uppsala, Sweden.

Abstract

This paper presents a method for monitoring the sludge profiles of a secondary settler using a Gaussian Mixture Model (GMM). A GMM is a parametric probability density function represented as a weighted sum of Gaussian components densities. To illustrate this method, the current approach is applied using real data from a sensor measuring the sludge concentration as a function of the settler level at a wastewater treatment plant (WWTP) in Bromma, Sweden. Results suggest that the GMM approach is a feasible method for monitoring and detecting possible disturbances of the process and fault situations such as sensor clogging. This approach can be a valuable tool for monitoring processes with a repetitive profile.

Keywords: signal monitoring, fault detection, clarifier, sludge profile

1 Introduction

The effluent water quality and efficient operation of resources are important aspects considered in the operation of a wastewater treatment plant (WWTP). Process monitoring and detection of abnormal conditions are crucial tasks, since they can help to improve the process performance (Olsson et al., 2014).

The sedimentation is an important process that determines the performance of the activated sludge process (ASP). The sedimentation is given by a secondary settler tank (SST), also called clarifier, which use gravity to separate the sludge (biomass) component from the treated water (liquid). Different approaches for predicting the SST behavior includes one, two or three-dimensional dynamic models. However, the prediction of the concentration profiles is still far from satisfactory (Li and Stenstrom, 2014), which makes the SST monitoring a complex task. Some examples of methods applied to monitor SSTs include image analysis (Grijspeerdt and Verstraete, 1997) and model-based approaches (Traoré et al., 2006; Yoo et al., 2002).

In the last two decades, a research field called *Machine Learning* has gained especial attention. The main scope with Machine Learning is to develop methods that can automatically detect patterns in data (learning), and then to use the uncovered patters to predict future data (Murphy, 2012). There are many different approaches in machine learning including decision trees, data clustering, neural networks, Gaussian process regression, Gaussian mixture models.

The authors proposed in Zambrano et al. (2015) an approach for monitoring a SST using Gaussian Process Regression (GPR), giving useful information about the status of the settler. GPR is a non-parametric regression method where data prediction is given as a probability density function. Hence, the predicted value comes with a variance estimate, interpreted as an uncertainty of the prediction. The method is thoroughly described by, for example, Rasmussen and Williams (2005) and Murphy (2012), and has gained large interest within the machine learning community for applications such as fault detection of environmental signals (Osborne et al., 2012), signal prediction (Grbić et al., 2013a,b) and control of bioreactors (Kocijan and Hvala, 2013).

In this work, we propose an alternative method for monitoring the process presented by Zambrano et al. (2015) based on a Gaussian Mixture Model (GMM). GMM is a parametric probability model for density estimation using a mixture of Gaussian distributions (Bishop, 2007). In this way, the GMM can describe a set of data using the combination of Gaussian distributions. Diverse applications of GMM can also be found in literature, for example in sensor monitoring (Zhu et al., 2014), fault detection and diagnosis (Yu, 2012).

The paper is organized as follows. First, a general introduction to GMM is presented, including a fault detection criteria based on the GMM formulation. Then, the problem of monitoring a secondary settler is presented as case study. Next, results and discussions are presented. Finally, some conclusions are drawn.

2 Materials and Methods

This section first presents the basics of Gaussian Mixture Models (GMM). Further, a GMM-based fault detection criteria is defined.

2.1 Gaussian Mixture Models

Assume we have a data vector x with N independent observations from a certain process. In a GMM, the total distribution of data is modeled as a sum (or mixture) of several Gaussian distributions with mean μ_k and covariance matrix σ_k. Hence, the model can be expressed as

(Murphy, 2012)

$$p(\mathrm{x}_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathrm{x}_i|\mu_k, \sigma_k),\ i = 1, ..., N \tag{1}$$

where each Gaussian distribution is denoted by $\mathcal{N}(\mathrm{x}_i|\mu_k, \sigma_k)$. The expression (1) is a combination of K Gaussian distributions, since we are taking a weighted sum. The mixing weights π_k must satisfy $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. The resulting function $p(\mathrm{x}_i)$ is a probability density function (pdf) from observing the data x_i.

When the value of K-groups is specified, the GMM parameters π_k, μ_k and σ_k can be inferred by using the iterative Expectation-Maximization (EM) algorithm applied to Gaussian Mixtures (Murphy, 2012), which can be summarized in Algorithm 1.

Algorithm 1 EM for Gaussian mixtures

1: Initialize $\mu_k^1, \sigma_k^1, \pi_k^1$ and set $i = 1$.
2: **while** not converged **do**
3: Compute $\gamma(z_{nk})$. ▷ Expectation step
4: Compute $\mu_k^{i+1}; \pi_k^{i+1}; N_k; \sigma_k^{i+1}$. ▷ Maximization step
5: $i \leftarrow i + 1$.
6: **end while**

The expressions used in Algorithm 1 are

$$\gamma(z_{nk}) = \frac{\pi_k^i \mathcal{N}(\mathrm{x}_n|\mu_k^i, \sigma_k^i)}{\sum_{j=1}^{K} \pi_j^i \mathcal{N}(\mathrm{x}_n|\mu_j^i, \sigma_j^i)}, n = 1, ..., N; k = 1, ..., K \tag{2}$$

$$\mu_k^{i+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathrm{x}_n, \tag{3}$$

$$\pi_k^{i+1} = \frac{N_k}{N}, \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk}), \tag{4}$$

$$\sigma_k^{i+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathrm{x}_n - \mu_k^{i+1}\right) \left(\mathrm{x}_n - \mu_k^{i+1}\right)^T. \tag{5}$$

One way to assign a value for K is using the silhouette criterion, see details in Rousseeuw (1987). The silhouette value S estimates how similar samples are in one cluster to samples in another cluster. S ranges from -1 (data misclassified) to $+1$ (data well-clustered), where S close to zero means that the clusters are indistinguishable.

2.2 GMM based fault detection criteria

When implementing a GMM to a group of data, the main idea is to compute a residual r so to monitor and decide between normal and abnormal profiles in the process. We assume that r belongs to one out of two different hypothesis: H_0 and H_1. Hence, the problem can be expressed by the classical binary hypothesis testing problem

$$\begin{aligned} H_0 &: r \leq h \\ H_1 &: r > h \end{aligned} \tag{6}$$

where H_0 refers to the non-faulty (normal) condition hypothesis, H_1 refers to the faulty (abnormal) condition hypothesis, and h is a predefined threshold. The aim is to decide if the system has changed between H_0 and H_1 when changes in the dynamic of the process are presented. It is assumed that H_0 and H_1 are equally likely.

For monitoring a group of profile data, each of them with N samples, we propose a GMM based residual r as detailed in Algorithm 2.

Algorithm 2 GMM-based residual calculation

1: Collect a group of M-profiles in non-faulty conditions.
2: Set K and compute the iterative EM algorithm (see Algorithm 1) to get π_k, μ_k, σ_k.
3: **while** monitoring a new profile **do**
4: **for** every profile **do**
5:

$$r = \frac{1}{p(\mathrm{x}; \pi_{1:K}, \mu_{1:K}, \sigma_{1:K})}, \tag{7}$$

where

$$p(\mathrm{x}; \pi_{1:K}, \mu_{1:K}, \sigma_{1:K}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathrm{x}_n|\mu_k, \sigma_k). \tag{8}$$

6: **end for**
7: **end while**

As given by expression (6), a fault is decided if $r > h$, where the threshold $h = \max\{r\}|_{t \in H_0}$ is the maximum r obtained during the evaluation of the non-faulty profiles. Hence, the non-faulty profile with data far from the rest of profiles will determine the value for h.

Note from expression (7) in Algorithm 2 that, the farther the new profile data is from the non-faulty data, the lowest the $p(\mathrm{x})$ and the larger the residual r will be.

3 Case Study: Monitoring a Secondary Settler

The present approach is tested using real data from a sensor installed in a secondary settler at Bromma WWTP in Stockholm, Sweden. The sensor measures the suspended solids (SS) concentration as a function of the settler level. The sensor goes from top to bottom of the setter, passing through the clarification and the thickening zone, and measuring the level [m] and the SS concentration [g/L], as shown in Figure 1(a). The profile obtained is called *sludge profile*. A typical sludge profile is shown in Figure 1(b).

Note in Figure 1(a) that we indicate a sludge blanket level, at which there is a jump from lower (less than 0.5 g/L) to higher (above 1 g/L) SS concentration, see Figure 1(b).

The sensor works discontinuously, which means that a new sludge profile is automatically measured after a certain period of time (in minutes). The collected data can be

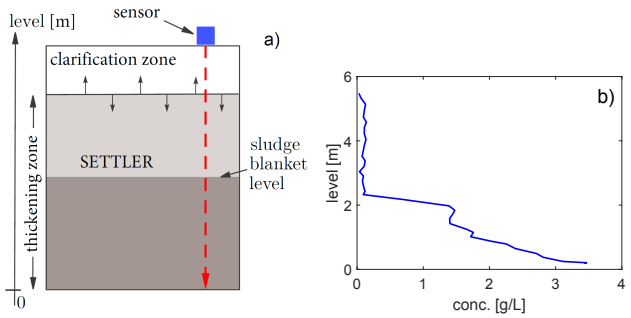

Figure 1. (a) Experiment setup; (b) Typical sludge profile plotted as level vs. SS concentration.

affected by different factors including: changes in the return and/or excess of sludge flow rates, sludge scape, large variations in the influent flow and composition and sensor clogging.

As part of the experiment, two additional measurements were recorded: the level at which the SS concentration is equal to 0.5 g/L (called *fluff level*) and equal to 2.5 g/L (called *sludge level*). We will refer to these levels during the results and discussions of the experiment.

4 Results

Figure 2(a) shows $M = 15$ sludge profiles in non-faulty conditions used for calculating the GMM. Figure 2(b) shows the non-faulty sludge profiles plotted using dots. The highest silhouette value obtained was $S = 0.77$ with $K = 3$, which means that the optimal number of clusters is three, as shown in Figure 2(c). Figure 2(d) shows the contours of the probability density function of the GMM.

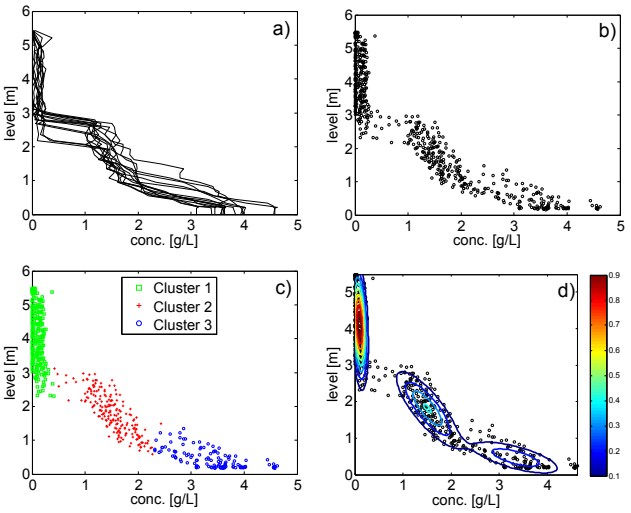

Figure 2. (a) Sludge profiles used to get the GMM; (b) Sludge profile data in (a) plotted using dots; (c) Clusters of the data in (b); (d) Contours of the GMM pdf, color scale indicates the value of the pdf contours.

The GMM parameters π_k, μ_k, and σ_k obtained for the data in Figure 2 are shown in Table 1. There we denote $\mathrm{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$, where $x_1 = \{\text{SS conc.}\}$ and $x_2 = \{\text{level}\}$. Then μ_k and σ_k are

$$\mu_k = \begin{bmatrix} \text{mean}(x_1) \\ \text{mean}(x_2) \end{bmatrix}, \quad \sigma_k = \begin{bmatrix} \text{cov}(x_1,x_1) & \text{cov}(x_1,x_2) \\ \text{cov}(x_2,x_1) & \text{cov}(x_2,x_2) \end{bmatrix},$$

where $k = 1,2,3$ refer to Cluster $1,2,3$, respectively, as shown in Figure 2(c)-(d).

Table 1. GMM parameters

k	Weight (π_k)	Mean (μ_k)	Covariance (σ_k)
1	0.43	$\begin{bmatrix} 0.09 \\ 4.11 \end{bmatrix}$	$\begin{bmatrix} 0.0074 & -0.0223 \\ -0.0223 & 0.7084 \end{bmatrix}$
2	0.34	$\begin{bmatrix} 1.50 \\ 1.82 \end{bmatrix}$	$\begin{bmatrix} 0.1446 & -0.1840 \\ -0.1840 & 0.3550 \end{bmatrix}$
3	0.23	$\begin{bmatrix} 3.34 \\ 0.47 \end{bmatrix}$	$\begin{bmatrix} 0.3612 & -0.1208 \\ -0.1208 & 0.0866 \end{bmatrix}$

The monitoring of the settler was carried out in several trials. As illustration, we present one trial which consisted of 33 days of monitoring, where a new sludge profile was collected every 15 minutes, giving a total of 3168 sludge profiles. In order to see the evolution of the sludge profiles during time, they are shown after 10, 20 and 30 days of running the experiment, as shown in Figure 3(a)-(c), respectively.

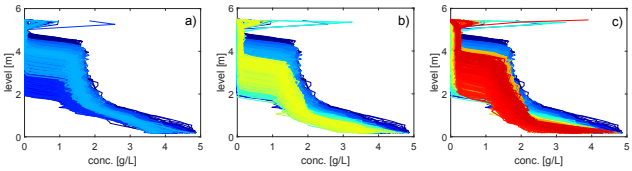

Figure 3. Total of sludge profiles during SST monitoring after: (a) 10 days; (b) 20 days; (c) 30 days.

Figure 4 shows the evolution of the fluff and sludge level, as well as the residual r. The residual r is colored from dark blue (beginning of experiment) to dark red (end of experiment), which correspond to the same range of colors assigned to the sludge profiles shown in Figure 3.

5 Discussions

As mentioned in the case study, a typical sludge profile has an abrupt change in the SS concentration around the sludge blanket level, see the profiles in Figure 2(a). This jump in the SS concentration was captured by the GMM, which classifies the data points before the jump as Cluster 1, and data points after the jump as Cluster 2, as shown in Figure 2(c). Also note that the data points with levels close to zero (bottom of the settler) and with high SS concentration were classified as Cluster 3.

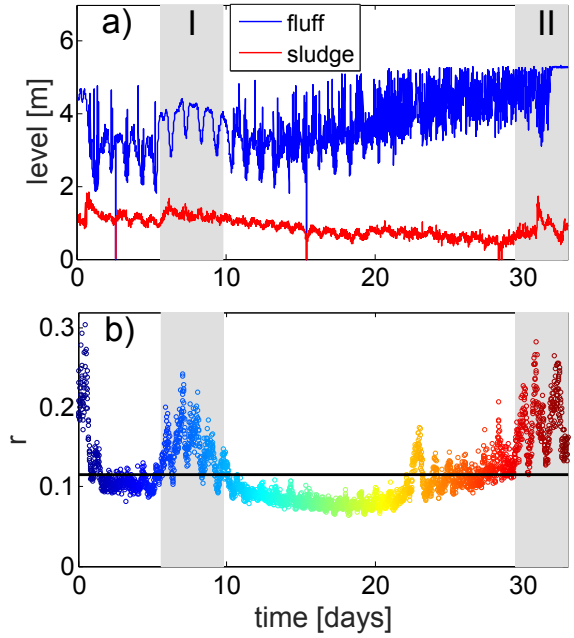

Figure 4. (a) Fluff level (blue line) and sludge level (red line); (b) Residual r (colored dots) and threshold h (black horizontal line). Gray zones refer to Period I and II, see details in Section 5.

From the total of profiles collected during the experiment, we highlight 2 groups of profiles marked as Period I and II in Figure 4. Period I refers to large variations in the influent flow rate, causing fluctuations in the sludge blanket, this effect can also be seen in the oscillatory variation of the fluff level (see Figure 4(a)). The sludge profiles of this period are shown in Figure 5(a). Note in this Figure that several data points at concentrations between 1 and 2 g/L are located far from the pdf contours with high values obtained from non-faulty data, which results in large values for r.

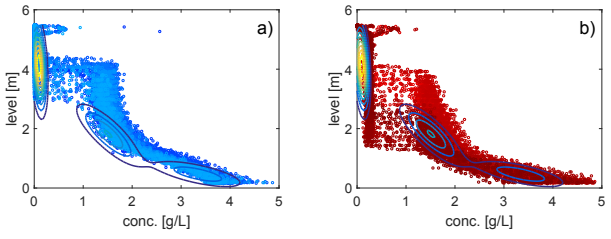

Figure 5. Group of sludge profiles for periods indicated in Figure 4. (a) Period I; (b) Period II. The plots include the contours of the probability density function shown in Figure 2(b).

Another type of events was related to sensor clogging, which began to be detected in profiles during Period II. This clogging event was confirmed by in-situ ocular inspection of the sensor and the existence of floating sludge at the surface level of the settler, promoting sludge escape. Figure 5(b) shows the sludge profiles of this Period. Note in this Figure that several data points are located far from the pdf contours with high values obtained from non-faulty data, particularly at concentrations below 0.5 g/L and between 1 and 2 g/L, which results in large values for r, sometimes even larger than those obtained in Period I.

Note that the data from both periods include outliers. Outliers are defined as sharp changes in the measured values between two successive data. For our case study, outliers in the sludge profiles mean that the measured data is far from the contours obtained with the non-faulty profiles (cf. Figure 2(d)). If there are several outliers in a given sludge profile, it will result in a large value for r. In this study, data correction from outliers was not part of the work. For a process with several events of outliers, the profiles reconstruction could be given by relocating the outliers using the GMM pdf.

Missing data is another possible situation when monitoring profiles. This is, when the amount of data in a given profile is incomplete. In the same way as in the case of outliers, the profiles reconstruction could be given by assigning the missing data using the GMM pdf.

Observe that collecting data from two sensors measuring the same process, the total set of data from each sensor will be different, resulting that each sensor will have a unique probability density function. This means that the present methodology has an important advantage, since is not just applied to specific sensors or processes but to sensors or processes from diverse areas.

A possible application for the current approach is to use the residual value r as a tool for a control action. In this way, it would be possible to formulate different control strategies based on, for example, changes in the recycle flow rate of the WWTP, in order to keep the new sludge profiles as similar as possible to the non-faulty profiles.

6 Conclusions

A GMM-based approach for monitoring and fault detection of the sludge profiles in a SST working in a WWTP has been proposed. Using a set of non-faulty profiles, the aim was to obtain a non-faulty region (SS concentrations, SST height) defined by a pdf via the GMM method. This pdf is then used to evaluate new profiles and detect possible abnormal profiles. Results obtained with real data implementation suggest that this method could help to monitor the performance of a SST.

Acknowledgment

The authors acknowledge funding support under the European Union's Seventh Framework Programme managed by the Research Executive Agency (REA), Grant Agreement N.315145 (Diamond). Funding from Käppala Association, Syvab and Stockholm Water Company, Foundation for IVL Swedish Environmental Research Institute and the Swedish Water and Wastewater Association is gratefully acknowledged.

References

Christopher Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007. ISBN 0387310738.

Ratko Grbić, Dino Kurtagić, and Dražen Slišković. Stream water temperature prediction based on Gaussian process regression. *Expert Systems with Applications*, 40(18):7407–7414, 2013a. doi:10.1016/j.eswa.2013.06.077.

Ratko Grbić, Dražen Slišković, and Petr Kadlec. Adaptive soft sensor for online prediction and process monitoring based on a mixture of Gaussian process models. *Computers & Chemical Engineering*, 58:84–97, 2013b. doi:10.1016/j.compchemeng.2013.06.014.

Koen Grijspeerdt and Willy Verstraete. Image analysis to estimate the settleability and concentration of activated sludge. *Water Research*, 31(5):1126–1134, 1997. doi:10.1016/s0043-1354(96)00350-8.

Juš Kocijan and N. Hvala. Sequencing batch-reactor control using Gaussian-process models. *Bioresource Technology*, 137:340–348, 2013. doi:10.1016/j.biortech.2013. 03.138.

Ben Li and M.K. Stenstrom. Research advances and challenges in one-dimensional modeling of secondary settling tanks – a critical review. *Water Research*, 65:40–63, 2014. doi:10.1016/j.watres.2014.07.007.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)*. The MIT Press, 2012. ISBN 0262018020.

G. Olsson, B. Carlsson, J. Comas, J. Copp, K. V. Gernaey, P. Ingildsen, U. Jeppsson, C. Kim, L. Rieger, I. Rodríguez-Roda, J.-P. Steyer, I. Takács, P. A. Vanrolleghem, A. Vargas, Z. Yuan, and L. Åmand. Instrumentation, control and automation in wastewater – from London 1973 to Narbonne 2013. *Water Science & Technology*, 69(7):1373, 2014. doi:10.2166/wst.2014.057.

Michael A. Osborne, Roman Garnett, Kevin Swersky, and Nando De Freitas. Prediction and fault detection of environmental signals with uncharacterised faults. In *Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, 2012.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series)*. The MIT Press, 2005. ISBN 026218253X.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi:10.1016/0377-0427(87)90125-7.

Adama Traoré, Stéphane Grieu, Frédérik Thiery, Monique Polit, and Jésus Colprim. Control of sludge height in a secondary settler using fuzzy algorithms. *Computers & Chemical Engineering*, 30(8):1235–1242, 2006. doi:10.1016/j.compchemeng.2006.02.020.

Chang Kyoo Yoo, Sang Wook Choi, and In-Beum Lee. Adaptive modeling and classification of the secondary settling tank. *Korean Journal of Chemical Engineering*, 19(3):377–382, 2002. doi:10.1007/bf02697143.

Jie Yu. A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. *Chemical Engineering Science*, 68(1): 506–519, 2012. doi:10.1016/j.ces.2011.10.011.

Jesús Zambrano, Oscar Samuelsson, Tatiana Chistiakova, Hongbin Liu, and Bengt Carlsson. Gaussian process regression for monitoring a secondary settler. In *2nd New Development in IT and Water*, Rotterdam, The Netherlands, 2015.

Hongyan Zhu, Shuo Chen, and Chongzhao Han. Fusion of gaussian mixture models for possible mismatches of sensor model. *Information Fusion*, 20:203–212, 2014. doi:10.1016/j.inffus.2014.02.002.