Supervisory control of wastewater treatment plants by combining principal component analysis and fuzzy c-means clustering

C. Rosen* and Z. Yuan**

*Dept of Industrial Electrical Engineering and Automation, Lund University, Box 118, 221 00 Lund, Sweden
**Advanced Wastewater Management Centre, University of Queensland, St Lucia, QLD 4072, Australia

Abstract: In this paper a methodology for integrated multivariate monitoring and control of biological wastewater treatment plants during extreme events is presented. To monitor the process, on-line dynamic principal component analysis (PCA) is performed on the process data to extract the principal components that represent the underlying mechanisms of the process. Fuzzy c-means (FCM) clustering is used to classify the operational state. Performing clustering on scores from PCA solves computational problems as well as increases robustness due to noise attenuation. The class-membership information from FCM is used to derive adequate control set points for the local control loops. The methodology is illustrated by a simulation study of a biological wastewater treatment plant on which disturbances of various types are imposed. The results show that the methodology can be used to determine and co-ordinate control actions in order to shift the control objective and improve the effluent quality.

Keywords: Fuzzy clustering; multivariate monitoring; PCA; set-point control; supervisory control; wastewater treatment

Introduction

In this paper an approach to an integrated multivariate monitoring and control system for wastewater treatment operation during extreme events (disturbances) is proposed. The method is based on multivariate statistics combined with clustering analysis to determine the operational state of the process. Information on the operational state is then used to determine appropriate set points for local control loops. The methodology is illustrated by a simulation study of a biological wastewater treatment plant on which disturbances of various types are imposed.

Automatic process control is today an important part of the operation of most biological wastewater treatment plants. The dissolved oxygen concentration is an example where automatic control has been successfully applied (Olsson and Newell, 1999). A local control system is typically concerned with a sub-process (called a unit process) of the whole system. Normally, these unit processes are controlled using a local feedback loop where the output of the process is measured and compared with a certain set point from which an appropriate control action is derived. Feed-forward control can also be utilised. Here, the control action is derived from a model describing the dependency of a process variable on a manipulative variable. An example is flow-rate proportional control of the return sludge flow rate.

The low-level control constituted by feedback and feed-forward control is usually sufficient under normal conditions when the characteristics of the influent wastewater are reasonably constant. However, as the operational conditions change, the control set points often have to be changed accordingly to obtain the desired operation. There are some different reasons why a supervisory control level is needed. Firstly, since the process is broken down into unit processes, there is a need to co-ordinate different control actions so that they do not have a counteractive effect. Secondly, a process may display non-linear behaviour...
when the operational conditions are far from the normal operating point, requiring changes to control set points. Thirdly, during extreme operational conditions such as hydraulic shocks or toxicity, the aim of the operation may shift significantly. Thus, a higher-level control system is needed to determine the control set points or control structure of the low-level control systems. This level is often referred to as the supervisory control level.

Supervisory control of wastewater treatment plants is typically performed by operators of the plants. That is, they take the responsibility for changing the control set points based on the information extracted from the on-line instrumentation data by an automatic monitoring algorithm or simply by directly visualising the process data. A natural step towards higher degree of automation in wastewater treatment plants is to close the loop by an algorithm, which automatically feeds the control set points to the local control loops. This has not been feasible in the past, but as the number of measured entities increase, the measurement quality improves and monitoring tools become more sophisticated, it is time for a discussion on how automatic supervisory control can be implemented in wastewater treatment.

Multivariate monitoring and classification

The objectives of process monitoring are to gather data and extract useful information from the measurements. Process data can be monitored using a wide range of different monitoring tools. Stochastic process control (SPC) (see e.g. Bissel, 1994) is a set of commonly used tools. In SPC, each variable is presented as a time series to the operator and control and alarm limits are used to define normal and abnormal operation. Multivariate techniques have become popular within many industrial fields as they can account for collective effects and reduce the dimensionality of the monitored data (see e.g. Wise and Gallagher, 1996). Principal component analysis (PCA) and other multivariate statistics (MVS) based techniques have proven useful for monitoring of wastewater treatment operation (Rosen and Olsson, 1998; Teppola et al., 1998).

The information obtained in the monitoring phase can be used to classify the current operational state. Knowledge based systems or expert systems have been used with varying success (Davies et al., 1996). Clustering techniques represent a different approach. Here, clustered data in the measurement space are said to represent similar process behaviour. Fuzzy c-means (FCM) clustering has been used to recognise clusters in wastewater treatment data (Marsili-Libelli and Müller, 1996). A combined approach of multivariate statistics and clustering to wastewater data monitoring can be found in Teppola et al. (1998). The approach presented here goes one step further. PCA and FCM are combined to determine the operational state of the process. The fuzzy information is then used to derive appropriate set points for local control loops in the process.

Principal component analysis

PCA can be described as a method to project a highly dimensional measurement space onto a space with significantly fewer dimensions. Often, several variables are highly correlated, since most variables only reflect a few underlying mechanisms that drive the process in different ways. This correlation is used in PCA to represent the underlying mechanisms as principal components (PCs). Let $\mathbf{X}$ be an autoscaled (i.e. mean-centred and scaled to unit variance) $[m \times n]$ matrix of measurement values for $n$ variables at $m$ number of samples defining a variable space of $r$ dimensions. The $r$-dimensional matrix $\mathbf{X}$ can be decomposed into a sum of the outer product of vectors $\mathbf{t}$ (scores) and $\mathbf{p}$ (loadings):

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_a\mathbf{p}_a^T + \mathbf{E} \quad \text{or} \quad \mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

where $\mathbf{E}$ is the residual matrix and $a \leq r$. If $a = r$ then $\mathbf{E} = 0$, as all variability is described.
However, if \( a < r \), i.e. less PCs than original variables are retained, then \( \mathbf{E} \) describes the variability not described by \( \mathbf{TP} \). Ideally, when \( a \) is chosen adequately, \( \mathbf{TP} \) describes the underlying mechanisms and \( \mathbf{E} \) represents the noise in matrix \( \mathbf{X} \). Often, in industrial systems \( a \ll r \), which implies a significant reduction of the number of dimensions (Wise and Gallagher, 1996).

PCA in its simplest form is a static modelling technique. However, there are ways to incorporate dynamics in the model, by including old measurement values in the analysis. The matrix \( \mathbf{X} \) is then constructed as:

\[
\dot{\mathbf{X}} = [\mathbf{X}_k \mathbf{X}_{k-l} \cdots \mathbf{X}_{k-j}]
\]

where \( \mathbf{X}_{k-l} \) means matrix \( \mathbf{X}_k \) translated \( l \) steps back in time. In this way, dynamic relationships between variables can be modelled.

The basic way of using PCA for monitoring involves identification of a model from data representing normal or desired operation. New data is then projected onto the model and the scores and/or the model residuals are then monitored as new samples are obtained:

\[
\dot{\mathbf{t}}_k = \mathbf{x}_k \mathbf{P}.
\]

It is important to note that new data are scaled in the same manner as data used for identification. Other available MVS-based methods include adaptive PCA (Wold, 1994; Dayal and MacGregor, 1997), principal component regression (PCR) and projection to latent structures (PLS) (see e.g. Wise and Gallagher 1996).

Fuzzy c-means clustering

Fuzzy c-means (FCM) clustering is a method that allows a certain instance to be a member of several classes at the same time, i.e. it is possible to be between two or more classes. Assuming that the cluster centres are known, the membership \( u \) to a certain cluster (class) \( i \) of an instance at time \( k \) can be calculated by:

\[
u_{k,i} = \left( \sum_{j=1}^{C} \left( \frac{d_{k,i}}{d_{k,j}} \right)^{2/(m-1)} \right)^{-1}
\]

where \( C \) is the number of classes and \( d_{k,i} \) and \( d_{k,j} \) are the distances from the instance to the centres of clusters \( i \) and \( j \), respectively. The parameter \( m \) determines the fuzziness of the classification. An \( m \) that is close to unity yields a crisp classification, whereas an increasing \( m \) makes the classification fuzzy. The distances can be calculated as:

\[
d_{k,i}^2 = (\mathbf{t}_k - \mathbf{v}_i)(\mathbf{t}_k - \mathbf{v}_i)^T
\]

where \( \mathbf{t}_k \) is the instance and \( \mathbf{v}_i \) is the cluster centre. The cluster centre can be determined manually. However, FCM is an unsupervised method, i.e. it can be used to find clusters in data. This is done iteratively for all instances of \( N \) using (6) to find a minimum:

\[
J_m(C,m) = \sum_{i=1}^{C} \sum_{k=1}^{N} (u_{k,i})^m d_{k,i}^2.
\]

The cluster centres are calculated from:

\[
\mathbf{v}_i = \frac{\sum_{k=1}^{N} (u_{k,i})^m \mathbf{t}_k}{\sum_{k=1}^{N} (u_{k,i})^m}.
\]
during the period of interest. Algorithms for adaptive FCM can be found in the literature (Marsili-Libelli and Müller, 1996; Teppola et al., 1998) and will not be discussed here.

Supervisory control by combining PCA and FCM

In this paper an approach to automatically determine controller set points (supervisory control) by means of combining PCA and FCM is proposed. This means that measurement data are projected as scores onto a smaller space defined by the principal components. Then, clustering analysis is carried out to locate the process in this space and, thus, determine the current operational state. This procedure has some advantages. Firstly, a reduced space through PCA implies reduced computational time for clustering analysis since the dimensionality of the problem is reduced. This is particularly beneficial for complex problems. Secondly, only variations represented by the model are projected as scores. Thus, measurement and process noise are, at least in the ideal case, not present in the scores. This means less misclassifications and more robust set-point determination. Thirdly, the convergence of the clustering algorithm is improved, as the scores are orthogonal. Teppola et al. (1998) have reported convergence problems in the clustering algorithm with highly collinear data. The supervisory control scheme is shown in Figure 1.

To translate the class membership to a control output or a set point, the classification needs to be defuzzified. The centre of area method is adopted here. The set point corresponding to a certain class is multiplied with the membership function for that class:

\[ sp_k = \frac{\sum_{i} u_{k,i}sp_i}{\sum_{i} u_{k,i}} \]  

Thus, the final set point, \( sp_k \), is calculated from the set points for all classes. Compared to the simple approach with which the class with the highest membership function is chosen, the centre of area approach utilises the fuzzy information on the class memberships. It is worth noting that, when FCM is used, the denominator in (8) is always unity. The set point for each class, \( sp_i \), can be predefined, which is the case in this work, or derived using models containing process knowledge.

Cases study: extreme event control

A simulation case study with application to wastewater treatment operation is reported here. In this study, a supervisory control component is designed to co-ordinate a couple of local control loops. The control objective is to protect the process under extreme conditions and, whenever possible, improve the effluent quality. The primary objectives of the study are to illustrate the proposed methodology and to evaluate its applicability to wastewater treatment operation.

Simulated plant

The simulated plant comprises two biological reactors and a secondary settler (Figure 2). Both reactors are aerated and, consequently, only carbon reduction and nitrification are of interest. The controlled variables considered in the design of the supervisory control system include dissolved oxygen (DO)-concentrations in both reactors, influent feed ratio between the two reactors and sludge-blanket level in the settler. The latter is controlled by manipulating the return-sludge flow rate. The IAWPRC Activated Sludge Model No 1...
(Henze et al., 1987) and a ten-layer one-dimensional settler model (Takács et al., 1991), are used to simulate the biological reactions and the settling process, respectively. Influent data developed by a working group on benchmarking of wastewater treatment plants within the European scientific research exchange programme COST 624 are used in the simulation (Vanhooren and Nguyen, 1996). The principal layout of the simulated plant is shown in Figure 2, together with important physical dimensions.

**Monitoring algorithm and local control systems**

The integrated control system consists of a number of different elements:

- monitoring of the influent water characteristics including flow rate, ammonia and suspended solids (SS) concentration measurements with a sampling rate of 4 h⁻¹;
- classification of the present operational state into one out of five different states;
- determination of set points using the newly proposed strategies (see below);
- a local control loop for the sludge-blanket level by means of return-sludge flow rate;
- local control loops for the DO-concentrations in the two reactors and a local control loop for the step-feed. The latter two control loops are assumed to be ideal, and hence their dynamics is not simulated.

In COST data there are three extreme events present: storm with sewer flush-out, i.e. high flow rate with high concentration of SS, storm with dilution and rain, also with diluted water. In addition to these disturbances an extreme ammonia load disturbance is included to mimic events that may occur in systems with anaerobic sludge treatment. Five operational states were defined accordingly: normal operation, storm with sewer flush out, storm, rain and high ammonia load.

A dynamic PCA model (see Eq. (2)), including one time-lagged value for each variable, is identified from influent data representing normal operation. From the PCA, three principal components are used to classify the operational state by means of FCM clustering. For
each defined disturbance, a training data set is constructed with similar disturbances to those of original COST data. Each training data set is projected onto the PCA model and two clusters are identified using FCM: normal and extreme event, resulting in a total of eight clusters. The clusters representing normal conditions coincide and, thus, five distinctive clusters are obtained. These five clusters with cluster centres are shown in Figure 3. Look-up tables with predefined set points are used to determine the set points for each operational state. Two values on m in (5) are evaluated: \( m=1.01 \), which corresponds to a crisp (non-fuzzy) classification and \( m=1.4 \), which yields a fuzzy classification.

Supervisory control strategies
A control strategy has been developed for each operational state. Each event strategy represents a shift in control objective from that of the normal operation.

1. **Normal operation strategy.** During normal operation, the sludge-blanket level is controlled to 0.68 m. DO-concentrations are 1.0 mg/l in both reactors and 100% of the influent flow is directed to the first reactor.

2. **Storm/flush-event strategy.** During flush-out, the sludge-blanket set point is raised to 2.0 m to lower the hydraulic load to the settler. Influent flow is directed to the first reactor.

3. **Storm strategy.** The strategy during storm events is to raise the sludge-blanket set point (1.5 m) for the same reason as mentioned above. The influent flow is redirected to the second reactor. In this way, sludge is accumulated in the first reactor with lower sludge load to the settler as a result. The idea is that this decreases the sludge loss to the effluent. The DO-concentration is lowered to 0.1 mg/l in the first reactor and raised to 2.0 mg/l in the second.

4. **Rain strategy.** The rain event strategy is similar to the storm strategy. The sludge-blanket set point is set to 1.0 m.

5. **Extreme ammonia load strategy.** The strategy during high ammonia load is to lower the sludge-blanket set point (0.2 m) in order to get a higher concentration of biomass in the reactors. The set point for the DO-concentrations is raised to 3.0 and 2.0 mg/l in the first and second reactor, respectively. 100% of the influent flow is directed to the first reactor.

A reference control case is defined to evaluate the performance of the controlled case. The return-sludge flow rate is proportionally controlled to 80% of influent flow rate. This rate yields the same sludge-blanket height as for the controlled case during normal operation. The excess sludge removal rate is controlled so that the reference and controlled case has the same sludge age. The DO-concentration is kept constant at 1 mg/l and no step feed is utilised.

Results and discussion
The results of the simulation study are discussed from two separate viewpoints: performance of the monitoring and control system and performance of the process.

Monitoring and control system
The main objective for the supervisory control system is to identify disturbances and to change the set points accordingly. Detection and an accurate classification are achieved during all disturbances. It is worth noting that the fuzzy classification \( (m=1.4) \) could detect a disturbance earlier and change the set points accordingly (Figure 4). Another advantage with the fuzzy classification is that the DO-set points were changed during ammonia peak loads, even though the operational state is considered normal (Figure 5). The fuzzy information also resulted in a smoother control with less saturation of control signals.
Figure 4: Detection and classification of the extreme ammonia load. Fuzzy (top) and crisp (bottom) classification with $m=1.4$ and $m=1.01$, respectively.

Figure 5: Fuzzy classification during normal operation (top). During peak load, the state is classified between class 1 and 5, resulting in a varying DO-concentration (bottom).

Figure 6: Set points for the sludge-blanket level at the start of the rain event using crisp and fuzzy classification.

Figure 7: Resulting return-sludge flow rate at the start of the rain event using crisp and fuzzy classification from Figure 6.
In Figure 6 the differences between the crisp and the fuzzy classification for generating set points for the sludge blanket level are shown. The gradual transition between two operational states implies a smoother set-point change and, consequently, a smoother change in the manipulated variable, the return-sludge flow rate (Figure 7). This has the advantage that hydraulic shocks are not introduced by the control system and that unnecessary wear on mechanical equipment is avoided.

Process performance
It can be seen from Figure 8 that the controlled (according to the proposed method) case yielded lower or as low effluent ammonia for the period of high ammonia load, but higher effluent concentrations during the periods of storms and rain. This is in compliance with the shift in the control objective—from carbon reduction and nitrification to prevention of sludge loss. This strategy shift is visible in the effluent SS concentrations (Figure 9). During the high ammonia load period, the effluent SS is in parity with the reference case, but during storm and rain periods, the effluent SS is decreased compared to the reference case.

It is important to note that the case study is carried out mainly to illustrate the applicability of the integrated monitoring and control strategy to wastewater treatment processes. There are a number of possible improvements. For instance, only influent data are used to monitor and classify the operational state. There is more information to gain from including process measurements in the analysis. Moreover, PCA-based monitoring as presented here
cannot handle changing process conditions. That is, the mean and variance of data have to be approximately constant over longer periods of time. Also, no consideration is taken to the fact that events occur in different time scales. There are solutions to these problems and they are discussed in Rosen and Lennox (2000). Another important improvement could be to incorporate the non-linear process behaviour in the monitoring model. This can be done by pre-processing of data or by using non-linear projection methods (Zhang et al., 1997).

An issue that has not been addressed here is the ratio between the sampling/updating-time of the low-level controllers and the supervisory control scheme. Supervisory control together with low-level controllers can be seen as cascade control. This means that the outer control loop (supervisory control) must have considerably slower dynamics than the inner loop (low-level control) to avoid oscillatory or unstable behaviour in the system. In this work, a sample/update rate of 4 h\(^{-1}\) was used for the supervisory control. This is perhaps too short in most real systems, especially if sludge dynamics are included. Therefore, in the general case the supervisory control may have to be separated into several levels with decreasing sampling/update rate depending on the dynamics of the controlled variables.

No strong conclusions with regard to the impact of the supervisory control system on the performance of the plant can be drawn from this preliminary study. However, the study indicates that the approach to supervisory control proposed in this work may be used to co-ordinate local control loops and to determine appropriate set points for the current operational state.

Conclusions
In this paper an approach to automatic supervisory control of wastewater treatment operation is proposed. By integrating on-line monitoring and control, appropriate low-level controller set point and structures for the current operational state of the process can be determined. The simulation study indicates that the proposed approach to automatic supervisory control is applicable to wastewater treatment operation. Principal component analysis (PCA) is a powerful method to extract relevant information from measurement data since it is capable of representing the underlying mechanisms by means of principal components (PCs). Fuzzy c-means (FCM) can be used to classify the operational state of the process and this is preferably done in the PC-space. A comparison of a fuzzy and crisp classification shows that fuzzy classification gives faster detection and smoother control than crisp classification.

References


