Membrane bioreactor fouling behaviour assessment through principal component analysis and fuzzy clustering

Thomas Maere\textsuperscript{a,}\textsuperscript{*}, Kris Villez\textsuperscript{b}, Stefano Marsili-Libelli\textsuperscript{c}, Wouter Naessens\textsuperscript{a}, Ingmar Nopens\textsuperscript{a}

\textsuperscript{a}BIOMATH, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000 Gent, Belgium
\textsuperscript{b}Laboratory of Intelligent Process Systems, School of Chemical Engineering, Purdue University, West Lafayette, IN, USA
\textsuperscript{c}Department of Systems and Computers, University of Florence, Via S. Marta 3, 50139 Florence, Italy

\textbf{Abstract}

Adequate membrane bioreactor operation requires frequent evaluation of the membrane state. A data-driven approach based on principal component analysis (PCA) and fuzzy clustering extracting the necessary monitoring information solely out of transmembrane pressure data was investigated for this purpose. Out of three tested PCA techniques the two functional methods proved useful to cope with noise and outliers as opposed to the common standard PCA, while all of them presented similar capabilities for revealing data trends and patterns. The expert functional PCA approach enabled linking the two major trends in the data to reversible fouling and irreversible fouling. The B-splines approach provided a more objective way for functional representation of the data set but its complexity did not appear justified by better results. The fuzzy clustering algorithm, applied after PCA, was successful in recognizing the data trends and placing the cluster centres in meaningful positions, as such supporting data analysis. However, the algorithm did not allow a correct classification of all data. Factor analysis was used instead, exploiting the linearity of the observed two dimensional trends, to completely split the reversible and irreversible fouling effects and classify the data in a more pragmatic approach. Overall, the tested techniques appeared useful and can serve as the basis for automatic membrane fouling monitoring and control.

© 2012 Elsevier Ltd. All rights reserved.

\section{1. Introduction}

Membrane bioreactors (MBRs) are nowadays used on a medium to large scale as state-of-the-art wastewater treatment for both industrial and municipal applications. The choice for MBR technology is especially made in the context of water reuse and specific site constraints. In other situations, MBRs are still seen as a high-cost option, mainly related to the technology's major drawback of membrane fouling. The latter influences the amount of membranes needed, as well as their life-time and operation (Judd and Judd, 2011). To prevent severe fouling problems, MBR installations are run...
conservatively. The usual approach consists of a fixed, experience-derived operational scheme of fouling suppression measures like membrane aeration, backwash and chemical cleaning in conjunction with routine analyses on membrane permeability to safeguard operation. However, such fixed schemes are never optimal, ignore the dynamic nature of fouling and lack the flexibility to cope with changing influent, biological and membrane conditions. This inevitably results in the waste of energy, permeate, chemicals and membrane life-time. Hence, significant potential for improving cost efficiency is available in the area of dynamic and online control (Busch and Marquardt, 2009; Drews, 2010).

Despite the latter, the number of contributions in literature dealing with dynamic and online fouling control is limited and full-scale implementations of such control systems are practically nonexistent (Ferrero et al., 2012). The problem lies mainly with the difficulty to obtain appropriate control inputs, i.e. variables that are unequivocally correlated to the fouling process (cause, state, effect) and measured or derived with a sufficient frequency to allow online control. An obvious choice would be to use transmembrane pressure (TMP), flux and permeability, since these variables are typically measured continuously in MBR plants and hold valuable information on the membrane state (Joss et al., 2009; Kraume and Drews, 2010). However, their application in control systems, either through mechanistic and semi-empirical models (e.g. Busch and Marquardt, 2009; Drews et al., 2009) or expert knowledge (e.g. Ferrero et al., 2011; Smith et al., 2006; Vargas et al., 2008), is hampered by the complexity of the process and strong coupling of different fouling phenomena.

An alternative approach is the use of designated monitoring instruments (e.g. Galinha et al., 2011; Huyskens et al., 2008; Mehrez et al., 2007) as illustrated by Huyskens et al. (2011) with a control scheme based on separate, in situ filtration test equipment to evaluate the mixed liquor's reversible and irreversible fouling propensity independently from membrane history and operating conditions. Although this approach certainly has its merits, such instruments are prone to error and require, apart from investment, additional maintenance and calibration. The representativeness of the obtained measurements has to be taken in consideration as well. The relatively low correlation that Huyskens et al. (2011) found between mixed liquor fouling propensity and online membrane permeability, for instance, suggests that control systems solely based on sludge quality and not accounting for membrane state and history, may not be adequate at all times.

In this study, a data-driven approach is proposed. It is based on principal component analysis (PCA) and fuzzy clustering (FC) and aims to extract the necessary control information from online TMP measurements, making it applicable to any MBR as this is a routine measurement. The strength of PCA and FC lies in their ability as data mining techniques to reveal and recognize patterns in vast amounts of data in the absence of exact process knowledge (Fu, 2011; Venkatasubramanian et al., 2003). As such, they seem genuinely suited as monitoring tools for membrane fouling and hence may lead to a control system capable of handling the complexity of the process without the need of additional equipment. The explicit use of PCA and FC for monitoring and controlling membrane fouling has not yet been reported in literature, but similar applications exist in various other fields, for instance for the monitoring and control of sequencing batch reactors and the detection of faults in conventional, continuous wastewater treatment plants (e.g. Aguado and Rosen, 2008; Marsili-Libelli, 2006; Villez et al., 2008). Hereby, especially the batch-wise PCA approaches appear interesting given the cyclic behaviour of the filtration process. However, the transfer of the reported methodologies to the MBR filtration process is not straightforward and an in-depth investigation is required to find the most adequate solution.

In what follows, the general PCA–FC concept is further introduced and three distinct algorithms, created to automatically infer the membrane state from TMP data, are presented. The specific features of these algorithms are discussed and their performance on a historical data set from a lab-scale MBR evaluated. Strong emphasis is placed on the algorithms' potential to serve as basis for a future real-time fouling control scheme.

## 2. Materials and methods

### 2.1. Data collection

The transmembrane pressure (TMP) data used in this study were collected between 18 November 2009 and 30 April 2010 from a lab-scale, side-stream MBR conceived for nutrient removal and fed with a synthetic wastewater resembling domestic sewage (108 l d⁻¹). The UCT-type (University of Cape Town) MBR was operated in quasi steady state with a hydraulic and sludge retention time of 6.4 h and 17 d, respectively. The temperature was controlled at 15 °C. A tubular Pentair Airlift X-Flow (The Netherlands) membrane module (material: polyvinylidene fluoride; pore size: 0.03 μm; tube diameter: 5.2 mm) was used for filtration and operated under a constant regime of 450 s filtration (31.8 l m⁻² h⁻¹), 18 s backwashing (106 l m⁻² h⁻¹) and 7 s relaxation (no flow). The mixed liquor suspended solids concentration in the membrane compartment was kept at 11 g l⁻¹ whereas it was 9 g l⁻¹ in the bioreactor. The membrane’s sludge and air crossflow velocities were set at 0.5 m s⁻¹ each. TMP measurements were obtained every second by means of three pressure sensors (Sensortechnics, Germany) and LabVIEW 7.1 (National Instruments, USA). More detailed information on the fully automated MBR setup and its influent can be found in Jiang et al. (2009).

An overview of the full TMP data set is given in Fig. 1, as well as a close-up of some exemplary filtration cycles. The membrane module was replaced three times by a new one, instead of chemically cleaning it. As such, four filtration periods exhibiting clean to severely fouled membrane behaviour can clearly be distinguished. Distinct events (peak values, recurrent zero values) are related to reactor maintenance and pump calibration. Sawtooth patterns, especially visible at the end of the first period, are related to operator intervention to prolong the membrane’s operation, i.e. a short series of backwashes at elevated membrane scouring air flows.
2.2. General algorithm scheme

The flowchart in Fig. 2 illustrates the employed monitoring concept and its algorithmic organization. The filtration cycles were first sorted according to their duration. The cycle end was detected by checking the backwash pressure drop $\Delta P$ (time $t_1$ – see Fig. 4). An identified cycle lasting 475 s (backwash, relaxation and filtration) was considered a potentially good cycle for assessing the membrane state. Each filtration cycle represented a single multivariate observation of the membrane state and was fed to the PCA as such (common PCA) or after being functionally approximated (functional PCA). The information content of the principal components to be retained was checked by reverse PCA prior to fuzzy clustering. The latter was used to describe the gradual transition from a clean membrane to a progressively fouled one. It was decided to have only three clusters to represent the main possible states and to limit the possible intermediate conditions. The inverse PCA transform was used to reconstruct the parameters in their original reference space after the clustering, enabling easy interpretation of the results. After proper training the combined PCA-FC algorithm could be used in real-time to infer the membrane state and decide on possible maintenance actions (indicated by dashed lines in Fig. 2). Hereby, new data are processed and confronted with the models calibrated on historical data. After classification these new data could be added to the historical data set for adaptive

Fig. 1 – Overview of the collected transmembrane pressure data (top) and a close-up of typical filtration cycles (bottom). The data set is subdivided in 4 periods, each ending with a membrane replacement.

Fig. 2 – General algorithm scheme starting from TMP data (top left corner) to the membrane assessment (bottom right corner). Solid and dashed lines are related to algorithm training and real-time assessment respectively.
model training. However, this real-time training, diagnosis and control aspect was not yet considered in current paper in view of an ad hoc expert system to be designed in future. The MATLAB R2010b (MathWorks, USA) platform was used for algorithm implementation and data processing.

2.3. Principal component analysis (PCA)

Segmentation of the continuous TMP data into consecutive filtration cycles led to a series of multivariate observations (assuming 475 separate TMP variables per cycle) constituting a two-dimensional data set \( (N \times 475) \), i.e. cycle number \( N \) versus TMP, on which common PCA, denoted as PCA from here on, was directly applied. PCA is widely used as a multivariate statistical technique to compress highly dimensional and correlated data into a few uncorrelated latent variables (linear combinations of original variables) that explain a maximum of variance (Jolliffe, 2002). These latent variables, called principal components (PC), represent the main underlying mechanisms that drive the process and can be used to visualize and interpret hidden phenomena in the data. Detailed information on the technique in the context of wastewater treatment can be found in Aguado and Rosen (2008), among others, and will not be repeated here. As all variables in the data set had the same unit (kPa), a classic covariance-based PCA was executed, meaning that the data were mean-centred but not scaled in any way prior to PCA. The number of PCs to be retained was based on a cumulative explained variance threshold of 95% and related analysis of the eigenvalue scree plot (Johnson and Wichern, 2002; Jolliffe, 2002). Parameters of the approximating functions. Advantages of this approach include that (1) the estimated functions typically express expert knowledge and the PCA model becomes easier to interpret, (2) the choice for a particular function allows to focus the PCA analysis on variations of interest and (3) the number of PCA parameters to be estimated is (typically) reduced because the number of parameters is generally lower than the number of variables in each time series, further leading to reduced uncertainty in the PCA model parameters. The analysis process can be reversed in the sense that the PCA model allows reconstruction of the parameters which in turn can be used for function evaluation to obtain filtered data (denoted as reverse PCA/preprocessing in Fig. 2). Two types of FPCA were studied in this paper. They are now further explained.

Expert-driven FPCA

In this first approach, 5 parameters \((+\Delta P, -\Delta P, a, b, S)\) were assumed to characterise a filtration cycle and estimated according to Fig. 4 for every cycle. The pressure at the end of filtration, denoted by \( +\Delta P \), gives an indication of fouling in general, while the negative pressure \( -\Delta P \) resembles the backwash phase (median value estimate) and thus by definition only irreversible fouling. During filtration, the TMP curve consists of an exponential part at the beginning and a subsequent increasing linear part. Hence, the parameters \( a \) and \( b \) of an exponential model and the slope \( S \) of a linear model were estimated using a least squared error curve fitting approach. Parameters \( a \) and \( b \), of which the latter is generally not used in literature, are likely influenced by both reversible and irreversible fouling, whereas \( S \) is supposedly related to reversible fouling only. Before PCA was carried out on the \( N \times 5 \) data set, the parameters were smoothed using a cubic spline filter to account for outliers. In contrast with common PCA, the data were standardized, i.e. mean-centered and scaled to unit variance, to remove artificial differences in importance caused by the differing parameter units.

B-splines FPCA

A second approach for FPCA was based on the fitting of a function guided to a lesser extent by first principles knowledge, i.e. a spline curve expressed as linear combination of B-splines (de Boor, 2001). In particular, cubic B-splines were

---

**Fig. 4** — Expert-driven FPCA cycle parameters for a typical filtration cycle beginning at \( t_1 \) when the pressure is reversed to backwash. Time \( t_1 \) is retrieved by comparing the TMP and its first and second derivative to specific thresholds (\( T_\text{h} \)).

\[
\begin{align*}
  t_1: \quad & \frac{d^2\text{TMP}}{dt^2} < T_\text{h} \land \frac{d\text{TMP}}{dt} < 0 < \frac{d^2\text{TMP}}{dt^2} < T_\text{h} < \text{TMP} > T_\text{h} \\
  +\Delta P: \quad & \text{TMP}(t) - \text{TMP}(t) \\
  -\Delta P: \quad & \text{TMP}(t) - \text{median}(\text{TMP}(t)) \mid t_1 < t < t_2 \\
  \text{TMP}(t) - \text{TMP}(t) = a \left( 1 - e^{-t/T_a} \right) \mid t_2 < t < t_4 \\
  \text{TMP}(t) - \text{TMP}(t) = b \left( t - t_d \right) + m \mid t_4 < t < t_5
\end{align*}
\]
chosen. These are very efficient for smooth functional approximation and consist of a collection of third order (cubic) polynomials with continuity constraints at fixed points called knots (Ramsay and Silverman, 2005). The knots in this study were placed at prescribed locations to best mimic the filtration cycle shape. Quadruple knots were used so to obtain step-like discontinuities in the overall fitted function, corresponding to the first drop in TMP (start backwash), first increase in TMP (start relaxation) and second increase in TMP (start filtration). Additional single knots were set by trial and error until a reasonable fit could be retrieved for the majority of the recorded cycles. Ultimately, a spline function supported by a basis of 48 B-splines was used, as shown in Fig. 5, resulting in a functionalised $N \times 48$ data set containing the respective linear combination coefficients to fit each filtration cycle. Ill-fitted cycles, identified through the root mean squared residual (RMSR) fit objective, were visually inspected and removed from further analysis when related to an anomaly in the data. The remaining cycles were processed by classic covariance-based PCA.

2.5 Fuzzy clustering

The information provided by the previous PCA methods can be used to classify the operational state of the filtration process. Clustering can assist in identifying the diagnostically significant regions in the principal component space (Aguado et al., 1998; Gustafson and Kessel, 1979). From all possible choices, fuzzy clustering was preferred for its ability to reflect the gradual changes of the fouling process in the partial degree of membership to the relevant clusters describing the membrane state. More specifically, the Gustafson-Kessel (GK) fuzzy clustering algorithm (Babuska et al., 1998; Gustafson and Kessel, 1979) was chosen for its flexible metric and ability to conform the clusters to the data. As any fuzzy clustering algorithm, it is based on the constrained minimization of the squared sum of distances between the data points representing the membrane cycles and each cluster mean (centroid), weighted by their corresponding fuzzy membership. A limit of the GK algorithm is that each cluster volume (i.e. the size of each cluster) is fixed a priori. To relax this constraint and achieve a better adaptation to the data, the cluster volumes were adapted at each iteration as in Krishnapuram and Kim (1999). The fuzzy exponent $m$, defining the sharpness of the partition, was determined ad hoc. The quality of the clustering was checked by the fuzzy normalized partitioning entropy ($H$) which indicates the residual uncertainty in the partition ($H = 0$: deterministic clustering, $H = \ln(k)$ with $k$ the number of clusters: most uncertain clustering) and the separation coefficient ($g$) defined as the maximum ratio between the sum of the cluster radii and the intercluster distance, i.e. representing the degree of overlap between adjacent clusters ($g < 1$: well separated, $g = 1$: tangent clusters, $g > 1$: overlapping clusters).

3. Results and discussion

3.1 Common PCA and FC

Common PCA was applied to the $N \times 475$ data set, $N$ equalling a total of 28,277 good length filtration cycles or 98.0% of all identified cycles (see Section 2.2 and 2.3). Two out of 475 possible PCs were retained, capturing respectively 96.1 and 3.4% of the variance in the data, making a combined explained variance of 99.5%. The loadings for PC1 and PC2, which reveal how the original variables are reflected in the PCs, are shown in Fig. 6. The first component (PC1) exhibits a clear resemblance to the general shape of the filtration cycle (Fig. 1), with a contrast between the backwash and filtration variables. High PC1 values would denote a high contrast, implying severe membrane fouling. PC2 accounts for phenomena that are not incorporated in PC1 and shows deviations in the backwash and, to a lesser extent, in the beginning and slope of the filtration section. High scores for PC2 would imply less negative backwash TMP values as determined by PC1 while the filtration values would remain unchanged, apart from a steeper beginning and more flat slope.

Fig. 7 presents the PC scores assigned to all individual filtration cycles. The different colours indicate membrane replacement (see Fig. 1). The score variations in time are not explicitly shown in this plot but are more clearly depicted in Fig. 9. Roughly, the initial scores for each membrane are located at the left, while final scores lie at the right (bottom or top). The results indicate two major trends: 1) scores shifting in a top-left to bottom-right direction; 2) scores moving in an almost perpendicular direction from the former. Membranes 2 and 3 exhibit solely the last behaviour, whereas membrane 4 solely the first and membrane 1 a combination of both. Referring to Fig. 6, the first behaviour can be related to more negative and positive TMP values during backwash and filtration respectively, while the second behaviour relates mainly to increasing filtration TMP values. All membranes were of the same type and the MBR operating conditions did not change during the studied period. Therefore, the observed
behaviour must somehow be related to changes in sludge characteristics leading to different fouling phenomena. A detailed investigation of this was beyond the scope of this paper.

As to the fuzzy clustering both $H(0.46)$ and $g(2.96)$ indicate a relatively uncertain partition and overlapping clusters. This is a consequence of having constrained the number of clusters to three and is also due to the nature of the data, gradually shifting from clean to fouled membrane filtration behaviour and therefore not creating clearly distinct clusters. Adjusting the fuzzy exponent $m$ changes the cluster shapes but the separation remains roughly the same. Though a non-optimal partition was produced, the three cluster centroid locations can be related to clean membrane filtration behaviour and both aforementioned fouling trends. In this sense the GK technique appears a useful tool for classifying the filtration data in conjunction with PCA decomposition, as was the aim (Fig. 2). On the other hand, with the obtained contour lines not all data can be correctly allocated. The 0.5 contours of the clean cluster (left), for instance, also reach out to the middle and lower right hand side of the figure. This could lead to a wrong classification of possible future data. Increasing the number of clusters did not improve the discriminatory ability of the partition. The issue may be related to the comprehensive nature of the GK algorithm requiring the ‘total membership’ assumption, i.e. that for each data point the sum of cluster memberships must be 1. Possibilistic clustering (Pal et al., 2005) could be used instead as it relaxes this constraint, but poses convergence and initialization problems and thus was not attempted here.

Fig. 8 illustrates the reconstruction of the filtration cycles by inverse PCA. The shown cycles are representative of the observed behavioural trends in Fig. 7 as they each pertain to differing clusters. The reconstruction proved satisfactory, although deviations at the beginning of the backwashing and relaxation phase are discernable for the clean membrane cycle.

The variation of the PC scores in time is shown in Fig. 9, together with the three cluster centroids (horizontal lines). The values for PC1 are generally rising except for the distinct sawtooth patterns related to operator intervention as mentioned in Section 2.1. The scores for PC2 are both rising (membranes 1, 2, 3) and descending (membranes 1, 4). This coincides with the earlier observed trends in Fig. 7. The

![Fig. 6](image_url) Loadings of the retained principal components for common PCA.

![Fig. 7](image_url) PCA score plot with fuzzy clustering centroids and contours representing cluster membership. The fuzzy exponent $m$ was set at 2.25.

![Fig. 8](image_url) Reconstruction of the first (top), median (middle) and last (bottom) filtration cycle after PCA, which are typical for respectively the left, top and bottom cluster in Fig. 7 with membership degrees above 0.75.
distinct outliers for membrane 4 are also due to operator intervention, i.e. manual deaeration of the membrane permeate lines during backwashing. It appears that PC1 gives an indication about the general severity of the fouling state of the membranes, whereas PC2, although accounting for a much smaller part of the variance in the data, enables the distinction between the two observed fouling phenomena by analysing the trends (rising, descending). Therefore, not just the score values, but also their trends are interesting features to further investigate and use for classification.

3.2. Expert-driven FPCA and FC

Parameterisation of the filtration cycles according to the scheme of Fig. 4 led to a $N \times 5$ data set upon which common PCA was applied. Two PCs were retained explaining respectively 81.0 and 14.9% of variance, as such reaching the targeted 95% limit. Retaining three PCs explained an additional 3.8% of variance, but did not substantially improve the results nor their interpretation (not shown). The loadings for PC1 and PC2 are shown in Fig. 10. All variables contribute almost equally to PC1, albeit for $\Delta P$ in an inverse way. Regarding PC2, $b$ and $\Delta P$ give the highest contribution whereas $a$, $+\Delta P$ and $S$ a smaller one. Here, the contribution of $S$ is reversed.

The scores are shown in Fig. 11 and exhibit the same major trends as in Fig. 7, albeit with some more variation. It is interesting to see that the expert FPCA discriminates between membranes 2 and 3. This was not the case for common PCA, but does make sense when looking at the raw data or the scores in Fig. 9 where different behaviour at the beginning of both membranes is apparent. The clustering yielded similar results as before with all 3 cluster centroids located in meaningful positions but the clustering contours not allowing a straightforward classification of all data.

The reconstruction of the 5 functional parameters through inverse PCA transformation, shown in Fig. 12, can be considered satisfactory. Including the third PC eliminated the observed deviations for $-\Delta P$ and $S$ but did not cause a major shift of the cluster centroids or scores and was thus not considered, as already stated. Looking at how each membrane is characterized by the selected parameters, some interesting observations can be made. The information contained in parameters $a$ and $+\Delta P$ appears very similar for all four membranes.
membranes, which was not expected. The latter parameter is influenced to full extent by both reversible and irreversible fouling and can thus be seen as an indicator of general fouling severity. However, the former, representing the membrane state at the beginning of the filtration stage, should by definition be less influenced by reversible fouling. Nonetheless, both parameters only differ slightly with regard to the cluster positions. For \( a \), the cluster centroids are evenly distributed over its range, while \( +\Delta P \) makes more distinction between cluster 1 (clean membrane behaviour) and clusters 2 and 3 (fouled membrane behaviour). The resemblance between \( +\Delta P \) and the PC1 scores for common PCA (Fig. 9) confirms the latter to indicate general fouling severity, as stated before. Parameter \( S \) shows the same clean/fouled distinction as \( +\Delta P \), but here the order of clusters 2 and 3 is reversed. Apart from this, apparent similarities between \( S \), which describes cake layer deposition and thus reversible fouling, and parameters \( +\Delta P \), \( a \) and \( b \) exist. It leads to think that these parameters are all influenced by both reversible and irreversible fouling and, as such, might be interchangeable to some extent. Parameter \( b \), related to the exponential transition at the beginning of filtration, presents a clearly deviating behaviour from earlier parameters for the 4th membrane and enables discriminating the third cluster from the other two. The second cluster can be differentiated from the others by \( -\Delta P \). In fact, \( -\Delta P \) fully determines this cluster and because this parameter is totally defined by irreversible fouling the second cluster must be as well. As a result, the third cluster can be linked to reversible fouling.

From all 5 parameters \( +\Delta P \) and \( -\Delta P \) seem the most informative, being unambiguously linked to fouling, i.e. representing respectively ‘overall fouling’ and ‘only irreversible fouling’. Nonetheless, it is advisable to include all parameters in the analysis to capture as much details as possible, certainly if a longer full-scale data set would be used with likely more variation in the data. If the information content for some parameters would be alike, the analysis will show.

Paramount for the expert FPCA algorithm, but in fact for all PCA techniques, is a sufficient data frequency. It should be high enough to capture all cycle features to allow parameter estimation, which should normally not pose a problem as pressure sensors can easily measure every second. On the other hand, not all types of membranes allow the estimation of \( -\Delta P \) as backwash is not always performed (i.e. flat sheet membranes). In this case, especially parameter \( b \) should be further investigated in terms of explaining fouling phenomena.

An overview of the expert FPCA scores in time is not provided as they greatly resemble those for common PCA. However, it was noticed that the expert FPCA was less influenced by outliers, as can also be seen in Fig. 12. This is largely due to the cubic spline filtering of the parameters, but the extraction of the parameters has a smoothing effect on its own as well, since their estimation is based on multiple data points.

3.3. B-splines FPCA and factor analysis

Different from previous methods, a check for outliers was performed before the actual PCA by analysing the computed RMSR values after function fitting. A reasonable fit was apparent for the majority of the cycles. Bad fits were characterised by high values in the RMSR cycles. With a RMSR value above 0.3 were visually inspected for anomalies. Of these 844 cycles, 52 cycles were found anomalous and could furthermore be split into a set of 47 and 5 cycles respectively. The first set was identified as having anomalous TMP values during backwashing, while the second contained cycles with anomalous TMP values during filtration. It shows that through functional representation, outliers can be detected before the actual PCA analysis and removed, as such not influencing further results.

The PCA model for the remaining (28,225) cycles was based on the first two principal components, individually capturing close to 85% and 14% of the total variance. The associated principal scores are not shown here, since they exhibited close to identical patterns as for common PCA (Fig. 7), except for the outliers. Instead of fuzzy clustering, an oblique rotation of the principal components was carried out so that the two observed linear trends, as discussed earlier, would align with the two axes of the figure. In addition, the centre of the model was shifted as such that zero values would correspond to clean membrane behaviour. The result of this rotation and shift on the scores can be seen in Fig. 13. Note that the rotation was obtained by trial and error and constrained to preserve the total variance in the two scores. In what follows, the rotated PCA model will be referred to as a factor model (Johnson and Wichern, 2002), with the rotated principal components as factors and the rotated principal scores as factor scores.

Fig. 14 shows the same factor scores as in Fig. 13, yet as a function of time. The corresponding factor loadings are
given in Fig. 15. As a result of the factor analysis, the two fouling mechanisms earlier discussed in relation to common PCA and expert FPCA are now completely separated from each other. This is especially visible in Fig. 14 as compared to Fig. 9. Factor 1 fully represents irreversible fouling as it totally resembles the $\Delta P$ values in Fig. 12, though reversed. Factor 2 represents only the reversible fouling. Interestingly, the operator interventions at the end of membrane 1 (sawtooth patterns due to extra backwashes and temporary elevated membrane aeration) were sufficient to remove all reversible fouling but did not have any effect on the values for factor 1. The irreversible fouling for membranes 2 and 3 was already high from the start, indicating sludge filterability issues. The onset of these issues seems indicated by the sudden and rapid increase of factor 2 scores already at the end of membrane 1. One can notice that the irreversible fouling for membranes 2 and 3 did not rise appreciably any further during operation.

Likely, the actual membrane was protected by the secondary membrane, i.e. the layer of irreversible and reversible fouling. Whereas reversible fouling can be counteracted by backwashes and elevated membrane aeration (membranes 1, 2, 3), for irreversible fouling (membranes 1, 4) the cultivation of an effective secondary (cake) layer might prove useful for extending the membranes’ operation time. This could be attained by aerating the membranes less, but also by longer filtration cycles for instance. For severe filterability issues different actions will be necessary, such as the addition of flocculants.

The rotation and complete separation of the observed trends makes that clustering is not needed in this case. In fact, simple expert knowledge-based limit values on the factor axes may now suffice for classification purposes, as indicated in Fig. 13, although additional trending information, i.e. the first derivative indicating speed of transition, would be useful as well. The sectioning in Fig. 13 was performed ad hoc and resulted in an upper and lower part indicating reversible and no-reversible fouling and the lower part was further split in an irreversible fouling and clean membrane behaviour section.

The loadings in Fig. 15 indicate that irreversible fouling (factor 1) has an influence on all cycle $\Delta P$ values, except for relaxation. This corroborates the earlier made remarks regarding the representativeness of the expert FPCA parameters $a$, $S$ and $\Delta P$ for reversible fouling. The factor 2 loadings only correct for deviations during filtration which further confirms its meaning as reversible fouling indicator.

### 3.4 Monitoring potential of the different PCA algorithms

The shown results indicate a large potential for all three PCA variants as a basis for an autonomous and effective
monitoring system. A particularly favourable aspect is that only TMP measurements, which are already available with most MBR SCADA systems, were needed to distinguish membrane fouling severeness and reversibility, information that is very important towards adequate membrane management and decisions on inter alia the membrane aeration rate, filtration cycle length, backwash flux, flocculant addition and chemical cleaning. As compared to routine analyses, the PCA approach can easily be automated and take long time spans into consideration. The former is needed for dynamic and real-time control while the latter can provide additional information on the system status, i.e. related to membrane history, for cost-efficient control. The PCA approach simply goes beyond what can be routinely checked by a plant operator in a given time frame.

Although all PCA variants were able to capture the main trends in the data, both FPCA approaches are preferred over the common PCA approach. The function fitting process suppresses the effect of noise in the data and enables the detection of outliers before the actual PCA analysis. The functional approach also allows the simultaneous analysis of different length filtration cycles, e.g. as the result from control actions, when they can be described by the same functional parameters. Downsides of FPCA are the choice of control actions, when they can be described by the same different length filtration cycles, e.g. as the result from functional approach also allows the simultaneous analysis of detection of outliers before the actual PCA analysis. The function fitting process trends in the data, both FPCA approaches are preferred over by a plant operator in a given time frame.

Pros and cons of the techniques were assessed through their addition and chemical cleaning. As compared to routine aeration rate, filtration cycle length, backwash flux, flocculant and the spline basis for each determined beforehand. The two major linear trends were observed in the data which, through expert FPCA, could be linked to reversible and irreversible fouling. The subsequent fuzzy clustering was capable of detecting these trends and could place the cluster centres in meaningful positions, which assisted in data interpretation for condition monitoring. The discriminatory ability of the GK algorithm was, however, not sufficient to correctly classify all data and needs further testing on a larger experimental database.

Control actions such as an adapted filtration cycle regime or membrane aeration rate can be based on the classification status of newly measured data, i.e. clean, reversible or irreversible fouling. However, as illustrated throughout this paper, the processed data also contain additional temporal information that should be incorporated in any decision regarding control. When the membrane is identified as clean but shows a clear trend towards reversible or irreversible fouling, one can intuitively sense that a control action would be needed as well. Apart from the trend direction, also the speed of transition should be looked into. The statistical properties of PCA could be helpful in determining the significance of observed trends and definitely deserve further research.

4. Conclusions

- The combined use of principal component analysis and fuzzy clustering to extract the necessary membrane fouling control information for a membrane bioreactor solely from transmembrane pressure measurements was investigated.
- The three tested PCA techniques exhibited similar capacities to uncover trends and patterns, but the two functional approaches are preferred over common PCA as they allow a better handling of noise and outliers. Expert FPCA appears the most suited for future work as it enabled a comprehensive interpretation of the data and also is less complex than B-splines FPCA. The latter, however, does provide a more objective way for functional representation of the data compared to the expert approach.
- Two major linear trends were observed in the data which, through expert FPCA, could be linked to reversible and irreversible fouling. The subsequent fuzzy clustering was capable of detecting these trends and could place the cluster centres in meaningful positions, which assisted in data interpretation for condition monitoring. The discriminatory ability of the GK algorithm was, however, not sufficient to correctly classify all data and needs further testing on a larger experimental database.
- Factor analysis proved a more practical approach than fuzzy clustering for the data in this study. It utilized the linear nature of the observed trends and appointed each trend to a separate axis. As such, limit values on the axes sufficed for delineating data classes. Trend transition rates were not accounted for by either approach but should be incorporated.
- Overall, the employed monitoring approach seems promising. On one hand, the TMP data appear to contain the necessary information for membrane management. On the other hand, the described techniques provide means to efficiently extract the information in an automated way and use it for control purposes.
Acknowledgements

Thomas Maere is supported by the Institute for Encouragement of Innovation by means of Science and Technology in Flanders (IWT).

REFERENCES


