Data-driven fault detection methods for detecting small-magnitude faults in anaerobic digestion process

Pezhman Kazemi, Jaume Giralt, Christophe Bengoa and Jean-Philippe Steyer

ABSTRACT

Early detection of small-magnitude faults in anaerobic digestion (AD) processes is a mandatory step for preventing serious consequence in the future. Since volatile fatty acids (VFA) accumulation is widely suggested as a process health indicator, a VFA soft-sensor was developed based on support vector machine (SVM) and used for generating the residuals by comparing real and predicted VFA. The estimated residual signal was applied to univariate statistical control charts such as cumulative sum (CUSUM) and square prediction error (SPE) to detect the faults. A principal component analysis (PCA) model was also developed for comparison with the aforementioned approach. The proposed framework showed excellent performance for detecting small-magnitude faults in the state parameters of AD processes.

Key words | anaerobic digestion, BSM2, data-driven, fault detection, support vector machine

Pezhman Kazemi (corresponding author)
Jaume Giralt
Christophe Bengoa
Universitat Rovira i Virgili,
Departament d'Enginyeria Química, Avda. Paisos Catalans, 26, 43007 Tarragona,
Spain
E-mail: *pezhman.kazemi@urv.cat*

Jean-Philippe Steyer
LBE, Univ Montpellier, INRA,
102 avenue des Etangs, 11100 Narbonne,
France

INTRODUCTION

In recent decades, the anaerobic digestion (AD) process has shown great potential for treating organic wastes, due to its high efficiency in reducing the chemical oxygen demand (COD) and biogas production (Wellinger *et al.* 2013). AD is a complex process in which the organic matter is decomposed by anaerobic bacteria. The whole biological degradation consists of four steps: hydrolysis, acidogenesis, acetogenesis, and methanogenesis. In the first step, the complex organic compounds are converted into soluble monomers and then these monomers are converted to volatile fatty acids (VFAs), CO_2 and hydrogen by the following acidogenesis and acetogenesis steps. In the last step, the methanogenic bacteria convert CO_2, hydrogen, and acetate into biogas, a mix of CO_2 and CH_4. Different types of anaerobic digesters have been developed and implemented for the treatment of waste and energy production around the world (Sun *et al.* 2016). Although AD technology has already reached its maturity, failures and low performances are still common at full scale due to the lack of reliable and robust fault detection (FD) systems. Because of the AD process complexity, if failures happen, extreme effort and time are required to return the process into its normal operation. Therefore, early detection of abnormal conditions that may cause failure is imperative to prevent losses (Sánchez-Fernández *et al.* 2018).

Many measurements such as pH, VFAs, biogas production, COD, alkalinity, ammonia concentration, and so on exist for monitoring anaerobic digesters. Each of these parameters gives different information about the state of health of AD. Among these parameters, pH and biogas production show the overall AD plant health; therefore, these parameters are not suitable for early detection of possible faults. For instance, decreases in gas production or pH are signs of process instability that has already occurred. Other parameters such as alkalinity, VFAs, COD and ammonia concentration can indicate process imbalance beforehand, but they do not provide direct information regarding the exact cause of the process imbalance. Therefore, to identify the exact cause of imbalance in the process, further analysis of other parameters is needed (Weiland 2008).

The main goal of FD frameworks is to find the abnormal events. Abnormal events are non-common events when the process deviates significantly from its normal operation. Generally, the FD frameworks are classified into three groups: model-based, knowledge-based and data-driven based approaches. In model-based approaches, the precise mathematical model of the process should be developed. However, due to the complexity of biochemical processes, obtaining the first principle models is a very difficult task.

doi: 10.2166/wst.2020.026

On the other hand, knowledge-based approaches are based on sets of rules and process behavior information that are extracted by experienced plant operators. The accuracy of this method depends on the operator and engineer's knowledge about the processes. However, obtaining this deep knowledge is always time consuming and very difficult for complex processes. To the contrary, data-driven approaches are merely designed based on the historical and online data without any need to develop a mathematical model or intervention of human knowledge. These methods are very beneficial when obtaining process knowledge is not easy in practice. One of the main disadvantages of data-driven approaches is that they cannot be developed prior to the design stage, due to the lack of real process data. Different data-driven techniques such as neural network (NN) (Heo & Lee 2018), support vector machine (SVM) (Ni *et al.* 2011), principal component analysis (PCA) (Aguado & Rosen 2008; Haimi *et al.* 2016), Bayesian network (Amin *et al.* 2019) and Fisher discriminant analysis (FDA) (He *et al.* 2005) have been used for fault detection and diagnosis of different industrial processes. However, to the best of our knowledge, the data-driven techniques have been rarely used for FD in the AD process yet. Therefore, the main goal of this work is to develop a data-driven FD framework, capable of detecting random small magnitude faults in the state of the AD process. As mentioned earlier, VFAs concentration is considered as one of the most important variables in most AD monitoring strategies since its accumulation in the reactor can be interpreted as either an organic overload or an inhibition of the methanogenic bacteria due to the influence of other factors (Jimenez *et al.* 2015). The same approach could have been applied using, for example, ammonia instead of VFAs concentration. However, ammonia can only lead to an inhibition while VFAs concentration is a central and key variable that reflects the impact of many different factors of influence (e.g. inhibitions like the one that can produce too high an ammonia concentration but also organic overloading or the presence of a toxicant in the feeding line) and is thus a widely used health indicator for AD processes (Li *et al.* 2014; Wu *et al.* 2019). In the present study, VFA soft-sensors based on the SVM method have been developed and trained using simulated data from the benchmark simulation model No. 2 (BSM2) (Jeppsson *et al.* 2006). Residual signals, which are the difference between measured VFA and ones predicted by the soft-sensors, are then generated. The obtained residual signals can be used alone or combined with univariate Statistical Process Control (SPC) charts such as cumulative sum (CUSUM) charts to detect the faults when the residuals exceed a control limit. In addition to the VFA soft-sensor, a PCA algorithm is used for comparative analysis with the proposed FD method. In this case, the same input vector as the soft-sensors plus VFA measurements were used as inputs for the PCA method.

MATERIAL AND METHODS

Data collection and pre-processing

The first step in designing a data-driven soft-sensor is to obtain the process data. The design and evaluation of the SVM soft-sensor has been carried out using collected data from BSM2 (Nopens *et al.* 2010). The BSM2 is a simulation environment containing a plant layout, a simulation model, influent loads, simulation procedures and evaluation criteria elements in order to analyse and to evaluate the performances of wastewater treatment plants. Fourteen process variables obtained from BSM2 simulation and tested as possible input variables are listed in Table 1. All these variables are measured from the influent, effluent and gas line of AD every 15 min. The obtained simulated data from day 245 to 453 and day 453 to 474 were used as training and test sets for developing VFA soft-sensor respectively. It should be noted that due to the high number of data points (20,000), random sampling of the training set was performed and finally 2,000 data points were selected uniformly to reduce the computation time during model training. Moreover, before model construction and prediction, the weighted moving average (WMA) method was adopted to reduce noises in signals (Hota *et al.* 2017). In this work, 100 sampling data was chosen as the window length for all model construction. Furthermore, due to the different units of the measured variables, it is also crucial to scale and normalize them before developing the different models.

Table 1 | Obtained variables from BSM2

Variables	Unit	Variables	Unit
Effluent COD	$g\ m^{-3}$	CH_4 mol_fraction	–
Effluent alkalinity	$mol\ m^{-3}$	CO_2 mol_fraction	–
Influent TSS*	$g\ m^{-3}$	H_2 mol_fraction	–
Effluent TSS	$g\ m^{-3}$	Pressure	bar
Effluent pH	–	Effluent ammonia	$g\ m^{-3}$
Effluent BOD**	$g\ m^{-3}$	Influent flow	$m^3\ d^{-1}$
Gas flow	$m^3\ d^{-1}$		

*TSS: Total soluble solids the following; **BOD: Biological oxygen demand.

Feature selection

In order to develop a high accuracy VFA soft-sensor, the best combination of variables in Table 1 should be selected. Using many inputs for model development can increase the noise and directly affect the model accuracy. Therefore, by implementing feature selection, the most important variables are chosen and the dimension of the input vector for soft-sensor training is reduced. The other benefits of adapting the feature selection method are shorter time of training, ease of interpretation of models, reduction of overfitting and lower cost in data collection. To avoid this problem, the *fscaret* package of the R environment was used (Szlęk 2013; Eskandarian *et al.* 2017). The package was chosen because it uses many models to perform feature ranking, therefore the result is more reliable.

Support vector machine (SVM)

SVM is a popular machine learning method for regression and classification that was first introduced by Cortes and Vapnik (Cortes & Vapnik 1995). SVM has been suggested as an efficient method for solving a general-purpose problem. Generally, in SVM, the input data is mapped into a multi-dimensional feature space by using the kernel functions. Then, the linear regression is applied in the feature space. By applying this mapping procedure, the non-linear problem can be solved in a linear space. The most famous kernel functions are the polynomial kernel, the radial basis, the exponential radial basis and the multilayer perceptron kernel function (Liu & Lei 2018).

Principal component analysis (PCA)

Due to the large number of measurements, industrial processes often contain a huge amount of data. One of the common methods to deal with this problem is PCA. The PCA method linearly transforms correlated data into a set of linearly uncorrelated values called principal components (PC). Fault detection using PCA is performed by estimating the Hottelling's statics (T^2) and the square prediction error (SPE) (Sánchez-Fernández *et al.* 2018). T^2 index is the squared Mahalanobis distance of the retained PCs, designed to measure the variability of the mean and covariance within these PCs. The SPE statistic is the measure of the lack of fit for the PCA model (Jackson 1991). The training PCA model has been performed by using based command in the R environment.

Control charts

The VFA soft-sensor can be used in conjunction with the statistical control chart to detect abnormal behaviors. A control chart is a graphical technique wherein a value of specific statistics is presented over time and, for the normal operation of the process, the statistics must not pass a predetermined control limit. Within this fault detection framework, the simulated VFA is compared with the predicted VFA by means of the SVM soft-sensor; then, the residual is determined and used to develop different control charts for monitoring the AD process, including the following:

- Square prediction error (SPE) chart: this chart illustrates the squared residual error obtained by comparing the real values of VFA and the predicted ones obtained by SVM soft-sensor.
- Cumulative sum (CUSUM) chart: here, the method represents the cumulative addition of deviations in every observation. The CUSUM control charts have proved to be highly sensitive in detecting small magnitude faults. A CUSUM chart can be obtained by using the following equations (Khusna *et al.* 2018):

$$C_i^+ = \max\left[0, C_{i-1}^+ + x_i - (\mu_{i,c} + k)\right] \quad (1)$$

$$C_i^- = \max\left[0, C_{i-1}^- + (\mu_{i,c} + k) - x_i\right] \quad (2)$$

$$C_0^+ = C_0^- = 0 \quad C_i = \max\left[C_i^+, C_i^-\right] \quad (3)$$

  where k, $\mu_{i.c}$, C_i^+ and C_i^- are the slack variable, the mean under normal operation (from day 453 to 491) of the plant and the upper and the lower CUSUM statistics, respectively. The role of slack variables is to introduce robustness into the calculate statistics. If there is any fault in the system, the statistics (C_i) in Equation (3) increases, showing accumulations of small deviations in the mean. These accumulations are corrected by using the slack variable. Typically, k is selected as half of the standard deviation of the samples.
- T^2 Hotelling's chart: this chart is considered as a multivariate method and can be determined based on the Mahalanobis distance (MD) by the following equation (Roman 1994):

$$T^2 = (x_i - \bar{x})S^{-1}(x_i - \bar{x}) \quad (4)$$

Here x_i is a sample data, $\bar{x}$ is the mean vector and S is a sample covariance matrix. All the control charts have been estimated using the *qcc* package in R environment.

In addition to the VFA soft-sensor, a PCA method (Sánchez-Fernández *et al.* 2018) was also used for the comparison with the suggested methods. In this case, the same input vector as the SVM soft-sensor plus VFA were used as inputs for the PCA method. The general diagram of proposed SVM and PCA fault detection approaches is illustrated in Figure 1.

Bootstrap confidence limits

Commonly, the mentioned control charts assume that monitoring statistics follow a certain probability distribution. However, in the most nonlinear and complex situations, the process observations do not follow a specific probability distribution. For this reason, for calculating the control limit for all control charts, the bootstrapping approach was implemented. The function for calculating the bootstrap confidence limit has been written in R environment. Figure 2 shows the bootstrap approach to calculate control limits, and it is summarized as follows (Phaladiganon *et al.* 2011; Khusna *et al.* 2018):

1. Compute the M statistics (M could be C_i, T^2, and SPE) with n observations based on the normal process operation.
2. Considered $M_1^i, M_2^i, \ldots, M_n^i$ be a set of n M values from i^{th} bootstrap sample ($i = 1,\ldots, B$) randomly chosen from the initial M statistics with replacement. The value of B should be large enough (e.g. B > 1,000).
3. In each of B bootstrap samples, determine the $100 \times (1-\alpha)^{th}$ percentile value given a specified value α with a range between 0 and 1.
4. Determine the control limit by taking an average of B $100 \times (1-\alpha)^{th}$ percentile values ($\bar{M}_{100\times(1-\alpha)}$).
5. The estimated control limit can be used to monitor a new observation. Therefore, if the monitoring statistic of a new observation exceeds $\bar{H}_{100\times(1-\alpha)}$, it can be considered as a fault event. In the current work, confidence level α was equal to 0.99.

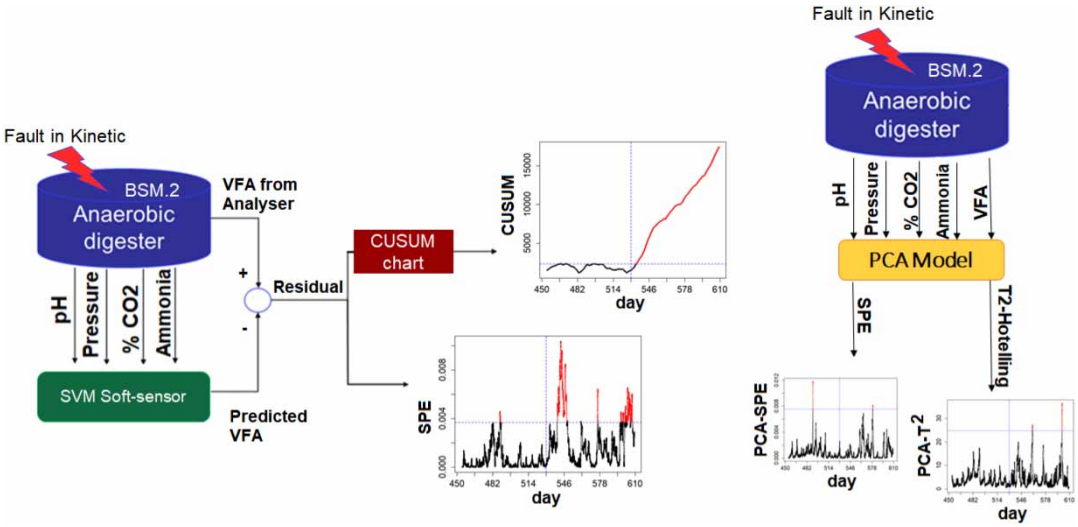

Figure 1 | General diagram of proposed SVM soft-sensor and PCA fault detection.

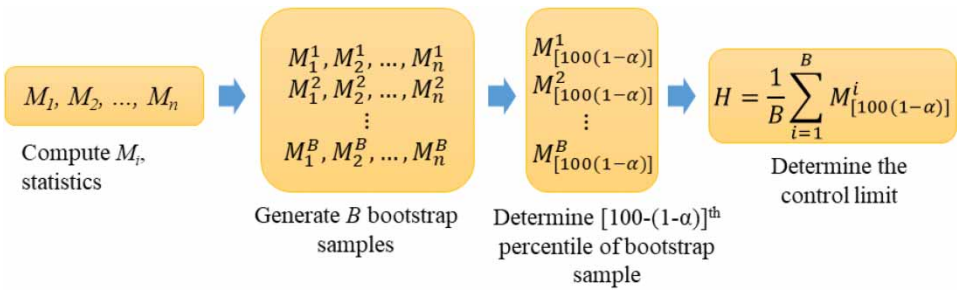

Figure 2 | Flow chart of the bootstrap method used for calculating the control limits.

Fault detection assessment

The performance of the proposed FD methods is measured in terms of precision, recall, and F1 scores given in Equations (5)–(7). Note that True positives (TP) are data points correctly labelled as faults. False positives (FP) refer to normal data points incorrectly labelled as faults. Finally, false negatives (FN) refer to faulty data points incorrectly labelled as normal.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$F1\ score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

The number of false detections and missed detections are captured by precision and recall respectively. The F1-Score is the harmonic mean of precision and recall. The higher values of these indicators show the higher performance of FD methods under evaluation. The delay, which indicates how slow the control charts are in detecting faults, is also calculated (Corominas *et al.* 2011).

RESULTS AND DISCUSSION

Linearity examination and feature ranking

Prior to developing the SVM soft-sensor, the linearity of the simulated data from BSM2 was investigated and it was found that the relationship between VFA and the other available parameters is highly linear in the relatively conservative AD process configuration included in BSM2. This linearity does not perfectly represent the situation in practice. Although this method works perfectly in linear situations, to increase the nonlinearity and make the objective more challenging, the input vector (feed) to AD, which consists of the concentration of inorganic nitrogen (S_{in}), composite (X_c), carbohydrate (X_{ch}) and also the feed flow rate, was manipulated.

In order to develop a high accuracy VFA soft-sensor, the best combination of available variables in Table 1 should be selected. Using many features for model development can increase the noise and directly affect the model accuracy. Therefore, feature ranking based on the *fscaret* package of the R environment (Szlęk 2013; R Core Team 2017) was performed on the available data to choose the best input subset. Before performing the feature ranking method, hard-to-measure parameters including COD, alkalinity and BOD were removed from the data set. The gas flow and CH_4 mole fraction were also removed due to their direct correlation with pressure and the CO_2 mole fraction respectively. It should be noted that the same results could be obtained by using gas flow and the CH_4 mole fraction instead of using pressure and the CO_2 mole fraction; therefore, the decision to eliminate correlated parameters can be made based on the simplicity and availability of measurements. The remaining parameters, listed in Table 1, were used as an input vector for the *fscaret* method. Figure 3 shows the importance of the variables on a scale from 0 to 100 obtained with the *fscaret* method for VFA prediction. In this figure, pH, ammonia concentration and pressure have the most influence on VFA.

Therefore, these important variables were chosen as the core subset and the other variables were added one by one to the SVM model based on their importance values.

The whole training procedure for the SVM soft-sensors was performed in the R environment by using the *Caret* package (Kuhn 2008). The tuning parameters of each model are estimated using the random hyper parameter *Search*, which is incorporated in this package. In this method, the tuning parameters of all models are randomly selected from the tuning space, which is defined beforehand. The number of randomly sampled values from the tuning space can be defined by the 'tuneLength' parameter of the *Caret* package (Bergstra *et al.* 2012). The test error is estimated for each subset after the models are trained. Table 2 shows the normalized root-mean-square error (NRMSE) of

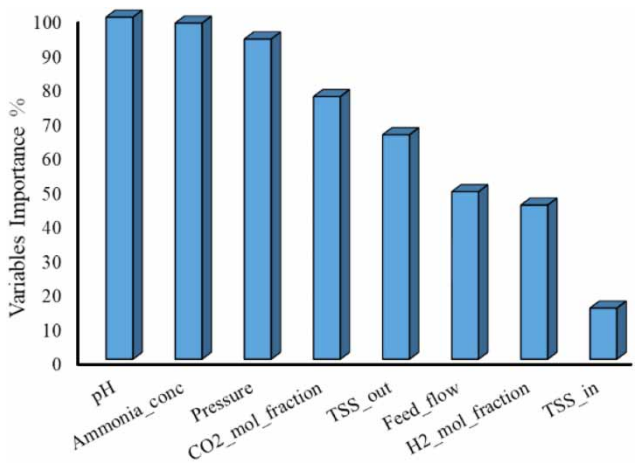

Figure 3 | Importance of variables on a scale from 0 to 100 obtained with the *fscaret* method.

Table 2 | Result of SVM soft-sensors trained using different input subsets

Inputs	R^{2a}	NRMSE[a]
pH + Ammonia_conc + pressure	0.813	0.182
pH + Ammonia_conc + pressure + CO_2_mol fraction	0.990	0.033
pH + Ammonia_conc + pressure + CO_2_mol_fraction + TSS_out	0.972	0.058
pH + Ammonia_conc + pressure + CO_2_mol_fraction + TSS_out + Flow	0.977	0.044
pH + Ammonia_conc + pressure + CO_2_mol_fraction + TSS_out + Flow + H_2_mol_fraction	0.971	0.049

[a]Estimated using test data set.

each subset based on the trained models. It can be seen that the second subset has the best NRMSE; therefore, this subset was selected for further development of the VFA soft-sensor. The kernel used for training SVM models was considered as the radial basis and best values of the two parameters C and γ for the final model (based on the second subset) were obtained equal to 245.88 and 0.0020 by the random hyper parameter *Search* respectively. The γ is a parameter for nonlinear hyperplanes and C is the penalty parameter of the error term (Cortes & Vapnik 1995).

Figure 4 represents the prediction result of the trained SVM soft-sensor on the test set. The NRMSE and R^2 were 0.04 and 0.98 respectively. As can be seen, the performance of VFA prediction by the developed soft-sensor is satisfactory and in the range of the best sensors available on the market for on-line VFA measurements.

Fault detection

To examine the detection accuracy of each method, the maximum uptake rate of acetate ($k_{m,ac} = 8\ day^{-1}$) in BSM2 was varied from ±5% to ±15% around its default value as

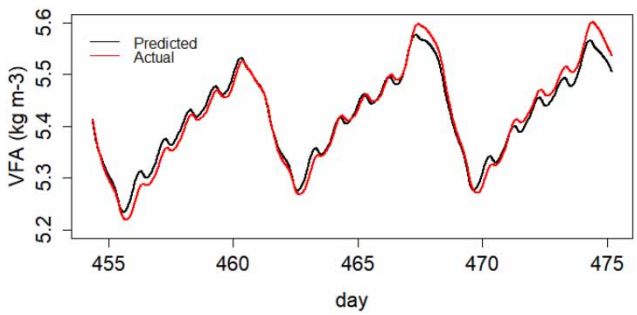

Figure 4 | Prediction result of SVM soft-sensor.

a simulated artificial fault. The variation of $k_{m,ac}$ can be similar to manipulation of the acetate concentration inside the digester. The fault starts from day 530 and lasts until the end of the simulation (609 days). The data from day 453 to 530 were defined as the normal data set and used for estimating the control limits of entire charts. The fault starts from day 530 and lasts until the end of the simulation (609 days). The fault detection performance of the developed SVM soft-sensor has been compared with the multivariate methods such as PCA and multivariate control chart T^2. The chart that has the highest F_1 score can be considered as the most precise one. The performances of the fault detection methods are presented in Table 3. In the PCA model, four PCs were selected to explain the 90% of variance. The control limit for all methods is obtained theoretically by using the bootstrap method with confidence level α equal to 0.99. Due to space limitation, only the performances for variation of +5% and +10% are shown.

It can be observed that the VFA-CUSUM chart has the highest F_1 score compared to the other charts, which means that it has better performance for detection of small magnitude faults. The precision for almost all models is in a good range, which means that all charts have a low false detection alarm rate. By comparing the delay of each chart, it can be concluded that due to the small-magnitude faults, all charts roughly have a high delay. The delay of VFA-CUSUM for +5% is the lowest among the other ones; however, when increasing the fault magnitude, the detection delay of the CUSUM method is also increased compared to the other charts. Although the delay of VFA-SPE and T^2 (5 inputs) charts for +10% is lower than the CUSUM chart, it is not clear whether this is a real fault or false detection due to its fluctuation around the confidence limit. However, due to the cumulative behavior of the CUSUM chart, it takes some time for the signal to overpass the threshold but as it has a positive trend (Figure 5(c)) it can definitely be considered as a fault. The PCA method has the lowest detection performance among all tested methods, which indicates that it is an improper choice for detecting small magnitude faults. Generally, it can be concluded that the CUSUM chart performs better on small magnitude faults; however, when increasing the faults' magnitude (> ± 15%), other charts can be more appropriate due to the inherent CUSUM delay. Therefore, to have a robust FD framework that is sensitive to both small and large magnitude faults, it would be better to use CUSUM jointly with the other charts.

As can be seen, the conventional SPE and T^2 control charts and the PCA method may not perform well for

Table 3 | Performance of each control chart for 5% and 10% variation of $k_{m,ac}$

	+ 5% $k_{m,ac}$				+ 10% $k_{m,ac}$			
Method and statistic	**Precision**	**Recall**	**F1 Score**	**Delay (day)**	**Precision**	**Recall**	**F1 Score**	**Delay (day)**
VFA-CUSUM	0.97	0.95	0.96	3.64	0.98	0.96	0.97	2.48
VFA-SPE	0.96	0.20	0.34	8.85	0.99	0.84	0.91	1.81
VFA-T^2 (5 inputs)	0.77	0.05	0.10	10.58	0.96	0.38	0.55	1.93
PCA-T^2 (5 inputs)	1	0.01	0.02	Not detected	1	0.1	0.19	6.01
PCA-SPE (5 inputs)	0.39	0	0	Not detected	0.83	0.03	0.06	Not detected

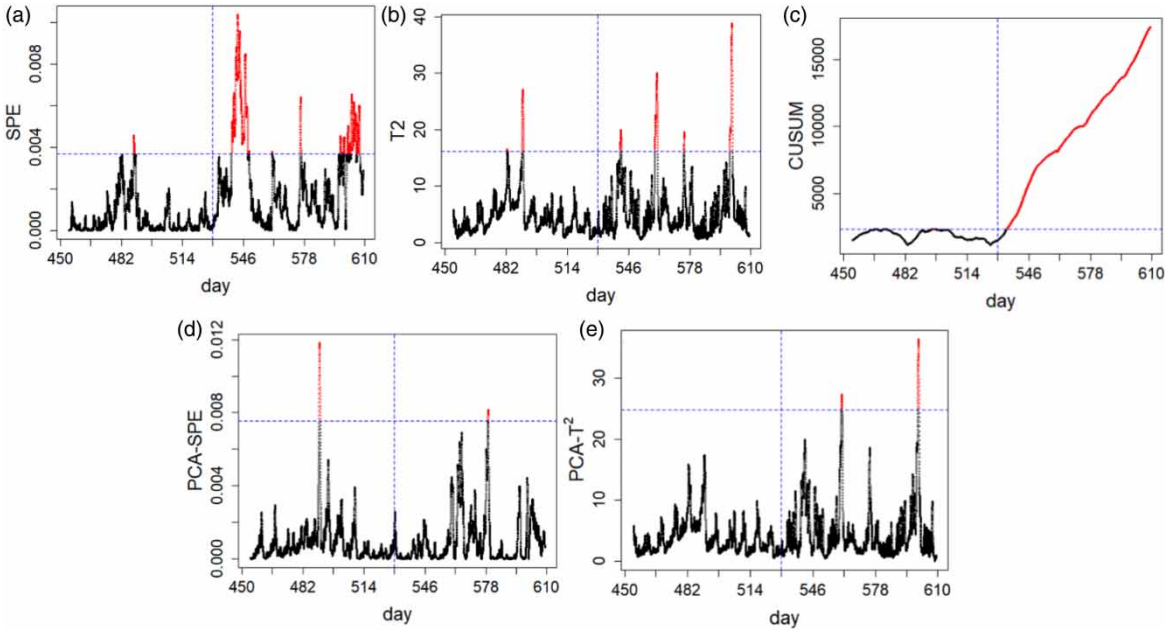

Figure 5 | Different control charts for +5% variation of $k_{m,ac}$. The horizontal and vertical blue lines show the obtained limit and the fault onset respectively. The full colour version of this figure is available in the online version of this paper, at http://dx.doi.org/10.2166/wst.2020.026.

small-magnitude faults. Although the CUSUM chart detects the fault with some delay, it still shows the best performance compared to the other charts.

CONCLUSIONS

This paper presented a combination of VFA soft-sensor and CUSUM chart as a dynamic and non-linear fault detection methodology for small-magnitude faults in the AD process. Prior to soft-sensor development, a feature selection method was used to find the most appropriate subset of measurements. Ammonia concentration, pH, pressure and CO_2 mole fraction were selected as the best subset of input variables. The performance of the suggested methodology has been compared with other conventional methods such as PCA and T^2 chart. CUSUM chart, based on the VFA soft-sensor residual, shows the best performance among the different approaches tested. Although it has shown some delay in detection, CUSUM drawbacks can be overlooked for small-magnitude faults, because of the high precision and recall values of the method, which can be interpreted as a low false alarm and a high detection rate respectively. It has also been concluded that the PCA method and T^2 chart are not robust for detection of faults with small magnitude. However, as the faults' magnitude increases, the performance of the CUSUM chart can also become low due to its high inherent delay. Therefore, to have an FD framework, which is robust in both high and low fault magnitude cases, it would be better to use CUSUM in combination with the other methods.

ACKNOWLEDGEMENTS

This project received support from the Ministerio de Economia, Industria y Competitividad, the Ministerio de Ciencia, Innovación y Universidades, the Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF), (CTM2015-67970-P, RTI2018-096467-B-I00). This article has been possible due to the support of the Universitat Rovira i Virgili (URV), (2017PFR-URV-B2-33, 2019OPEN) and Fundació Bancària 'la Caixa' (2017ARES-06). The authors' research group is recognized by the Comissionat per a Universitats i Recerca, DIUE de la Generalitat de Catalunya (2017 SGR 396). The authors are also very grateful to Dr. Ulf Jeppsson from the Lund University, Sweden, for his kindness in providing the BSM2 Matlab code.

REFERENCES

Aguado, D. & Rosen, C. 2008 Multivariate statistical monitoring of continuous wastewater treatment plants. *Engineering Applications of Artificial Intelligence* **21**, 1080–1091. Available from: www.elsevier.com/locate/engappai (accessed 21 November 2019).

Amin, M. T., Khan, F. & Imtiaz, S. 2019 Fault detection and pathway analysis using a dynamic Bayesian network. *Chemical Engineering Science* **195**, 777–790. Available from: https://www.sciencedirect.com/science/article/pii/S0009250918307371 (accessed 14 May 2019).

Bergstra, J., Ca, J. B. & Ca, Y. B. 2012 *Random Search for Hyper-Parameter Optimization Yoshua Bengio*. Available from: http://scikit-learn.sourceforge.net (accessed 15 May 2019).

Corominas, L., Villez, K., Aguado, D., Rieger, L., Rosén, C. & Vanrolleghem, P. A. 2011 Performance evaluation of fault detection methods for wastewater treatment processes. *Biotechnology and Bioengineering* **108** (2), 333–344. http://doi.wiley.com/10.1002/bit.22953 (accessed 21 November 2019).

Cortes, C. & Vapnik, V. 1995 Support-vector networks. *Machine Learning* **20** (3), 273–297. Available from: http://link.springer.com/10.1023/A:1022627411411 (accessed 9 April 2018).

Eskandarian, S., Bahrami, P. & Kazemi, P. 2017 A comprehensive data mining approach to estimate the rate of penetration: application of neural network, rule based models and feature ranking. *Journal of Petroleum Science and Engineering* **156** (2017), 605–615. Available from: http://dx.doi.org/10.1016/j.petrol.2017.06.039 (accessed 8 August 2018).

Haimi, H., Mulas, M., Corona, F., Marsili-Libelli, S., Lindell, P., Heinonen, M. & Vahala, R. 2016 Adaptive data-derived anomaly detection in the activated sludge process of a large-scale wastewater treatment plant. *Engineering Applications of Artificial Intelligence* **52**, 65–80. Available from: https://www.sciencedirect.com/science/article/pii/S0952197616300124 (accessed 26 March 2018).

He, Q. P., Qin, S. J. & Wang, J. 2005 A new fault diagnosis method using fault directions in Fisher discriminant analysis. *AIChE Journal* **51** (2), 555–571. http://doi.wiley.com/10.1002/aic.10325 (accessed 1 March 2019).

Heo, S. & Lee, J. H. 2018 Fault detection and classification using artificial neural networks. *IFAC-PapersOnLine* **51** (18), 470–475. Available from: https://www.sciencedirect.com/science/article/pii/S2405896318320664 (accessed 14 May 2019).

Hota, H. S., Handa, R. & Shrivas, A. K. 2017 *Time Series Data Prediction Using Sliding Window Based RBF Neural Network*. Available from: http://www.ripublication.com (accessed 1 August 2018).

Jackson, J. E. 1991 *A User's Guide to Principal Components*. Wiley, Hoboken, NJ.

Jeppsson, U., Rosen, C., Alex, J., Copp, J., Gernaey, K. V., Pons, M.-N. & Vanrolleghem, P. A. 2006 Towards a benchmark simulation model for plant-wide control strategy performance evaluation of WWTPs. *Water Science and Technology* **53** (1), 287–295. Available from: https://iwaponline.com/wst/article/53/1/287/12474/Towards-a-benchmark-simulation-model-for-plantwide (accessed 2 September 2018).

Jimenez, J., Latrille, E., Harmand, J., Robles, A., Ferrer, J., Gaida, D., Wolf, C., Mairet, F., Bernard, O., Alcaraz-Gonzalez, V., Mendez-Acosta, H., Zitomer, D., Totzke, D., Spanjers, H., Jacobi, F., Guwy, A., Dinsdale, R., Premier, G., Mazhegrane, S., Ruiz-Filippi, G., Seco, A., Ribeiro, T., Pauss, A. & Steyer, J.-P. 2015 Instrumentation and control of anaerobic digestion processes: a review and some research challenges. *Reviews in Environmental Science and Bio/Technology* **14** (4), 615–648. Available from: http://link.springer.com/10.1007/s11157-015-9382-6 (accessed 29 March 2018).

Khusna, H., Mashuri, M., Ahsan, M., Suhartono, S. & Prastyo, D. D. 2018 Bootstrap-based maximum multivariate CUSUM control chart. *Quality Technology & Quantitative Management* **17** (1), 1–23. Available from: https://www.tandfonline.com/doi/full/10.1080/16843703.2018.1535765 (accessed 2 February 2019).

Kuhn, M. 2008 Building predictive models in *R* using the caret package. *Journal of Statistical Software* **28** (5), 1–26. Available from: http://www.jstatsoft.org/v28/i05/ (accessed 11 April 2018).

Li, L., He, Q., Wei, Y., He, Q. & Peng, X. 2014 Early warning indicators for monitoring the process failure of anaerobic digestion system of food waste. *Bioresource Technology* **171**, 491–494.

Liu, L. & Lei, Y. 2018 An accurate ecological footprint analysis and prediction for Beijing based on SVM model. *Ecological Informatics* **44**, 33–42. Available from: https://www.sciencedirect.com/science/article/pii/S1574954117302972 (accessed 9 April 2018).

Ni, J., Zhang, C. & Yang, S. X. 2011 An adaptive approach based on KPCA and SVM for real-time fault diagnosis of HVCBs.

IEEE Transactions on Power Delivery **26** (3), 1960–1971. Available from: http://ieeexplore.ieee.org/document/5763739/ (accessed 1 March 2019).

Nopens, I., Benedetti, L., Jeppsson, U., Pons, M.-N., Alex, J., Copp, J. B., Gernaey, K. V., Rosen, C., Steyer, J.-P. & Vanrolleghem, P. A. 2010 Benchmark simulation model no. 2: finalisation of plant layout and default control strategy. *Water Science and Technology* **62** (9), 1967–1974. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21045320 (accessed 22 August 2018).

Phaladiganon, P., Kim, S. B., Chen, V. C. P., Baek, J.-G. & Park, S.-K. 2011 Bootstrap-based T^2 multivariate control charts. *Communications in Statistics – Simulation and Computation* **40** (5), 645–662. Available from: http://www.tandfonline.com/doi/abs/10.1080/03610918.2010.549989 (accessed 6 March 2019).

R Core Team 2017 *R: A Language and Environment for Statistical Computing*. Available from: https://www.r-project.org/

Roman, M. 1994 Practical decomposition method for T2 hotelling chart. *International Journal of Industrial Engineering: Theory, Applications and Practice* **20** (5–6). Available from: http://journals.sfu.ca/ijietap/index.php/ijie/article/view/684 (accessed 2 February 2019).

Sánchez-Fernández, A., Baldán, F. J., Sainz-Palmero, G. I., Benítez, J. M. & Fuente, M. J. 2018 Fault detection based on time series modeling and multivariate statistical process control. *Chemometrics and Intelligent Laboratory Systems* **182**, 57–69. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0169743918303459 (accessed 2 February 2019).

Sun, H., Wu, S. & Dong, R. 2016 Monitoring volatile fatty acids and carbonate alkalinity in anaerobic digestion: titration methodologies. *Chemical Engineering & Technology* **39** (4), 599–610. Available from: http://doi.wiley.com/10.1002/ceat.201500293 (accessed 2 February 2019).

Szlęk, J. 2013 The fscaret. Automated feature selection using variety of models provided by caret package. Available from: https://scholar.google.com/scholar_lookup?title=The fscaret %2C automated feature selection using variety of models provided by caret package&author=J. Szlęk&publication_year=2015 (accessed 15 August 2018).

Weiland, P. 2008 'Wichtige Messdaten für den Prozessablauf und Stand der Technik in der Praxis BT – Messen, Steuern, Regeln bei der Biogaserzeugung : 15. November 2007, Convention Center, Messe Hannover' in Gülzower Fachgespräche. Gülzow, Fachagentur Nachwachsende Rohstoffe, 17–31. Available from: https://www.openagrar.de/receive/timport_mods_00012091

Wellinger, A., Murphy, J. & Baxter, D. 2013 *The Biogas Handbook : Science, Production and Applications*. Available from: https://books.google.es/books/about/The_Biogas_Handbook.html?id=wUSAZwEACAAJ&redir_esc=y (accessed 17 February 2019).

Wu, Y., Kovalovszki, A., Pan, J., Lin, C., Liu, H., Duan, N. & Angelidaki, I. 2019 Early warning indicators for mesophilic anaerobic digestion of corn stalk: a combined experimental and simulation approach. *Biotechnol Biofuels* **12**, 106. https://doi.org/10.1186/s13068-019-1442-7 (accessed 10 January 2020).

First received 1 October 2019; accepted in revised form 20 January 2020. Available online 27 January 2020