Article

A statistically based fault detection and diagnosis approach for non-residential building water distribution systems.

Hafiz Hashim[1]*, Paraic Ryan[2, 3] and Eoghan Clifford[4]

[1]School of Engineering, National University of Ireland, Galway, Ireland.
h.hashim1@nuigalway.ie (H.H.), paraic.ryan@ucc.ie (P.R.), eoghan.clifford@nuigalway.ie (E.C.).
[2] Discipline of Civil, Structural and Environmental Engineering, School of Engineering, University College Cork, Ireland.
[3] MaREI Centre, Environmental Research Institute, University College Cork, Cork, Ireland.
[4] Informatics Research Unit for Sustainable Engineering, National University of Ireland, Galway, Ireland.
*Correspondence: h.hashim1@nuigalway.ie; Tel: 00 353 91 492219.

Abstract: Large non-residential buildings can contain complex and often inefficient water distribution systems. As requirements for water increase due to water scarcity and industrialization, it has become increasingly important to effectively detect and diagnose faults in water distribution systems in large buildings. In many cases, if water supply is not impacted, faults in water distribution systems can go unnoticed. This can lead to unnecessary increases in water usage and associated energy due to pumping, treating, and heating water. The majority of fault detection and diagnosis studies in the water sector are limited to municipal water supply and leakage detection. The application of detection and diagnosis for faults in building water networks remains largely unexplored and the ability to identify and distinguish between routine and non-routine water usage at this scale remains a challenge. This study, using case-study data, presents the application of principal component analysis and a multi-class support vector machine to detect and classify faults for non-residential building water networks. In the absence of a process model (which is typical for such water distribution systems), principal component analysis is proposed as a data-driven fault detection technique for building water distribution systems for the first time herein. Hotelling T^2-statistics and Q-statistics were employed to detect abnormality within incoming data, and a multi-class support vector machine was trained for fault classification. Despite the relatively limited training data available from the case-study (which would reflect the situation in many buildings), meaningful faults were detected, and the technique proved successful in discriminating between various types of faults in the water distribution system. The effectiveness of the proposed approach is compared to a univariate threshold technique by comparison of their respective performance in the detection of faults that occurred in the case-study site. The results demonstrate the promising capabilities of the proposed fault detection and diagnosis approach. Such a strategy could provide a robust methodology that can be applied to buildings to reduce inefficient water use, reducing their life-cycle carbon footprint.

Highlights:

- Principal component analysis and support vector machines were combined to build a fault detection and diagnosis method for non-residential building water networks.
- Data from a non-residential building was used to validate the developed fault detection and diagnosis model.
- Improved performance monitoring of non-residential water distribution systems can be achieved using limited process data.
- A comparative study illustrated that the proposed approach detected and diagnosed more actual system alarms than a conventional approach, while also utilizing less computational resources.

Keywords: Water distribution system, Non-routine events, Fault detection and diagnosis, Principal component analysis (PCA), Support vector machine (SVM), Performance monitoring.

1. Introduction:

In Europe, the non-residential sector consumes about 28% of total water withdrawn (European Commission, 2012). Thus, the efficient use of this water in such buildings is imperative. In certain areas of Europe, overall losses in water networks due to leakage are as high as 50% (Nowicki et al., 2012). For example, 41% of treated water is lost through damaged pipes in Ireland (Irish Water, 2019). It has been reported that water loss in buildings

is primarily linked to leakage within the water distribution systems (Datta & Sarkar, 2016; Skworcow et al., 2013) - in a recent study Sydney Water estimated that building leaks account for 28% of total water consumption (Sydney Water, 2011). It is thus vital to reduce the wastage of water through leakage and improve the performance of such water distribution systems (and thus reducing associated energy usage due to water pumping, treatment, etc.) This need has resulted in a growing focus on optimising water (and associated energy) usage across all building types using building management systems to detect problems (Sousa et al., 2019), particularly given the water scarcities impacting many countries today (Connor & Murphy, 2017).

In many sectors, fault detection and diagnosis systems have been developed to optimise system operation and to detect and diagnose system abnormalities. Examples include heating ventilation and air conditioning (HVAC) systems, water and wastewater treatment processes, automotive industries, chemical and petrochemical processes, wind farms and some manufacturing facilities (Hu et al., 2019; Xu et al., 2017; Zhang et al., 2017; Naderi & Khorasani, 2016; Zhou et al., 2015; Yu et al., 2014; Bruton et al., 2014; Zhang et al., 2017). For example, the analysis of energy consumption in buildings as a key driver of fault detection and diagnosis has been extensively studied (Hu et al., 2019; Burak Gunay et al., 2019; Agostino et al., 2017; Balaras et al., 2017; Stavset & Kauko, 2015). Previous studies have tended to describe one fault detection and diagnosis method at a time (either principal component analysis or support vector machines), often formulated as an optimization task. For instance, principal component analysis was used to compress the data or to identify the anomalies using detection indices, whereas the goal of using machine learning models (such as support vector machines) was to minimise the difference between the measurements taken on the system and predicted values from physics-based models (Grueiro et al., 2018; Escofet et al., 2016; Nasir et al., 2014). A significant challenge impacting the performance of these approaches is the time varying characteristics of real time processes (such as change in mean, variance over time, process noise, etc.) which can be a particular feature of water consumption.

Fault detection and diagnosis studies in the water sector have focused mainly on leak detection at a municipal water supply level (Seyoum et al., 2017). A range of detection approaches (such as leak-noise correlation, acoustic sensing, water balancing, pressure management, etc.) have been studied for these municipal systems (Robles et al., 2016; Escofet et al., 2016; Pérez et al., 2015; Makaya & Hensel, 2015). Although these approaches are considered the most accurate for leak detection, they are not suitable when it comes to large-scale building water distribution systems due to being relatively high cost, labour-intensive and time consuming. Other approaches have relied on hydraulic modelling techniques (such as transient analysis, pressure residual vector method, etc.) which have been validated using modelled or synthetic data. These approaches have focused on comparisons of measurements with predictions obtained from hydraulic models (Moors et al., 2018; Abdulshaheed et al., 2017; Alsaydalani, 2017; Soldevila et al., 2017; Perfido et al., 2016; Moser et al., 2015; Sedki & Ouazar, 2012; Mashford et al., 2012; Salam et al., 2015). Creating such models is often difficult and expensive due to the presence of non-stationarity and uncertainties in complex non-residential water distribution systems.

Overall, the implementation of fault detection and diagnosis for water distribution systems has received limited attention (Stetco et al., 2019; Geng et al., 2018). This is particularly the case for non-residential buildings (Adeyeye, 2014; Cosgrove et al., 2015) where it is acknowledged significant challenges remain in developing detection and diagnosis system for faults in building water distribution systems (Liu et al., 2016; Nezhad et al., 2014). Furthermore, fault detection and diagnosis methods must be cognisant with challenges associated with real data that comes from non-residential building water distribution systems (e.g. to distinguish between routine and non-routine water usage, etc.).

Currently, most building managers react to faults on an ad-hoc basis, responding to obvious faults and repairing infrastructure as required (Moser et al., 2015). This can result in low-level imperceptible faults persisting within a water distribution system, compromising water distribution system efficiency. The application of existing approaches developed for municipal water supply systems may be limited within non-residential buildings as key faults not only involve leaks but also comprise system related faults (such as equipment malfunctioning, operational errors, etc.) or non-routine events caused by inefficient water usage and infrequent sudden changes in demand (such as hosting a one-off event or increased production for a short period of time, taps left running, etc. - Figure 1). Furthermore, it is advantageous within buildings to distinguish between faults, inefficient water usage and unusual consumption which may or may not be the result of a fault.

Building water distribution system studies have not leveraged the more sophisticated fault detection and diagnosis approaches used in other sectors such as principal component analysis or support vector machines (Gharsellaoui et al., 2020; Zhao et al., 2019). Thus, the development and validation of robust systems that can detect system alarms, and distinguish between faults and non-routine events in non-residential water distribution systems has the potential to significantly increase the efficiency of water consumption (Pelz, 2003; Danacova et al., 2016). The present work aims to fill these gaps by developing decision models based on a real site data that can detect faults in non-residential buildings.

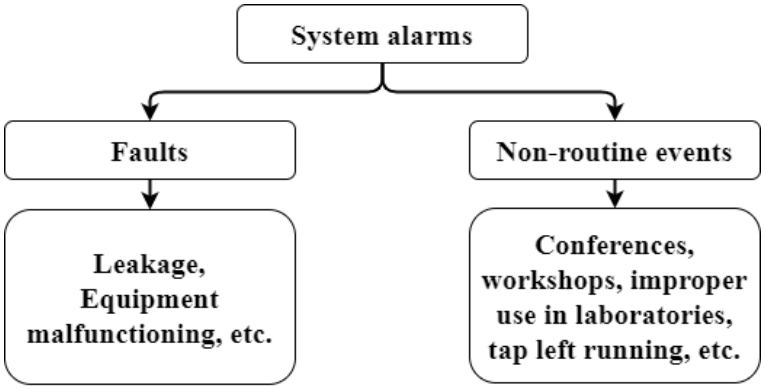

Figure 1: System alarm characterization.

In general, fault detection and diagnosis methods can be categorized mainly as: signal-based detection, and signal and multiple model-based detection (Figure 2) (Pelz, 2003; Venkatasubramanian et al., 2003a; Venkatasubramanian et al., 2003b).

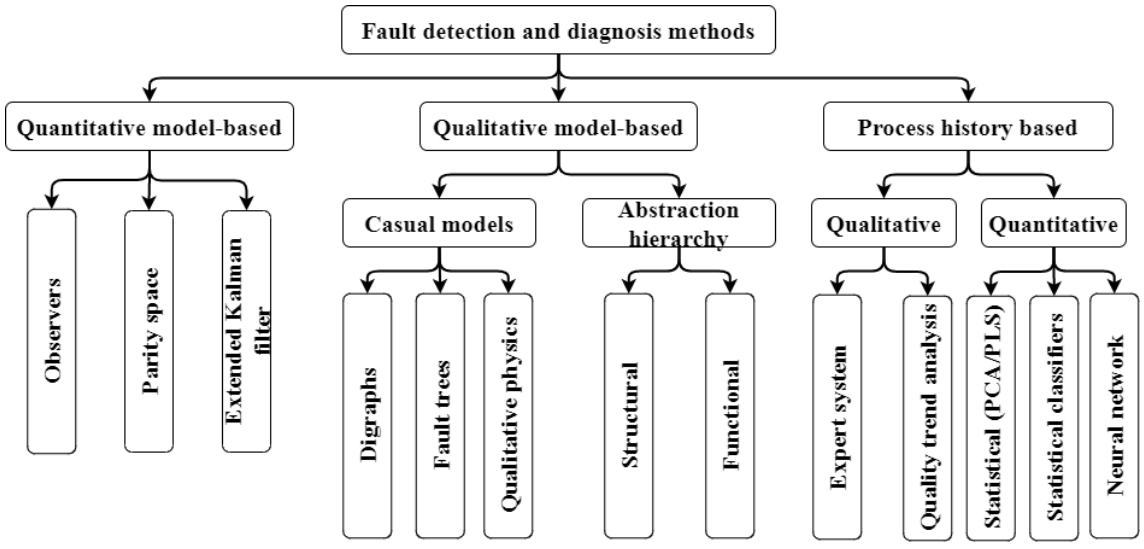

Figure 2: Fault detection and diagnosis methods. Adapted from (Venkatasubramanian et al., 2003; Pelz, 2003).

Basic fault detection systems comprise alarms which trigger when relatively high levels of water are being used (Mulligan et al., 2020; Clifford et al., 2018; Quevedo et al., 2014; Perfido et al., 2016) and provide high level statistics on water consumption. Typically, in such systems when water consumption exceeds a predefined threshold, an alarm is sent automatically. While the alarm may not be designed to detect excessive water consumption due to improper use (e.g. a tap left running) or other non-routine events, it can be adapted to do so. However, such systems are unable to identify complex patterns of water usage - such as distinguishing between actual faults or non-routine events. Additionally, such systems are embedded with alarm notification features but still require assistance from experts to choose thresholds for faults or non-routine events. More effective fault detection and diagnosis approaches must be capable of detecting inevitable deviations in water time series. Furthermore, non-residential water distribution system can feature non-stationarity statistical properties over time (e.g. water consumption can vary between seasons and depend on working hours, holiday periods, etc.) which can

make it challenging to detect faults (Prabuchandran et al., 2019). It can also be noted (e.g. Table 1) that a significant gap exists in published literature relating to fault detection and diagnosis methods that utilize real case-study data. Existing studies in the literature tend to focus on the use of generated or experimental data sets (Cody & Narasimhan, 2020; Grueiro et al., 2018; Moser et al., 2015; Villegas et al., 2010). The use of idealised data means the effect of random variations and outliers, which undoubtedly occur in industry are not taken into account in the assessment of proposed methodologies. This can lead to an over-estimation of the effectiveness of proposed approaches if applied to real case-studies.

Principal component analysis is a versatile and well-known technique for unsupervised data analysis, which projects data in a reduced space defined by orthogonal principal components (Izem et al., 2018; Wei Li et al., 2018). It has been adopted in various disciplines, such as in damage detection of civil structures, data compression, face recognition, image analysis, visualization as well as in fault detection in petroleum/chemical industries (Laory et al., 2011; Saitta et al., 2005; Posenato et al., 2008; Harrou et al., 2013; Abdi & Williams, 2010; Li et al., 2000). Principal component analysis (PCA) is a vector space transformation predominately used to reduce data dimensionality by extracting correlation between variables in sets of independent variables, that explain the trend of the process while optimising the variance of the original data in reduced number of dimensions (Garcia-Alvarez, 2014; Laory et al., 2011; Villegas et al., 2010). A PCA model leverages process history data and extracts a linear combination of variables explaining the major trend of the process being analysed (Ait-Izem et al., 2018).

To measure the variation of samples within the PCA model and detect the abnormality of the new incoming data, Hotelling T^2-statistics and Q-statistics (squared prediction error) have been applied in other engineering domains (Li et al., 2018; Horrigan et al., 2018; Zhang et al., 2017; Villegas et al., 2010). The T^2-statistics measure the variability of a score matrix and can detect abnormality within new data by comparing it to variation in the parameters defined by baseline condition. Q-statistics evaluate the variability of a matrix which is the projection of the original data onto a residual subspace (Benaicha et al., 2013; Ahmed et al., 2012; Zumoffen & Basualdo, 2008). PCA based methods have advantages over classical statistical methods (e.g. Shewhart chart, cumulative sum,, exponentially weighted moving average) which while accurate in many cases, have been shown to produce inaccurate results when introduced to multivariate systems (Maione et al., 2019; Xiao et al., 2017). Other challenges associated with the application of classical statistical approaches to building water distribution systems include the considerable variation in water consumption levels in buildings, and the possible mismatches between the proposed statistical model and the process being monitored (Chen, 2010).

In recent times, the use of support vector machines learning has increased relative to other methods (e.g. artificial neural network) when modelling the relationship between the input and the output in classification and regression problems in various engineering sectors (Xu et al., 2020; Kouziokas, 2020; Gautam et al., 2020; Cody, 2020; Akil et al., 2019; Grueiro et al., 2018; Liu et al., 2019; Samer et al., 2018). This is due to SVM's strong fault diagnosis ability in multi-class classification problems, high training efficiency, and its effectiveness when working with low-dimensional (small sample dataset) process data (Xu et al., 2017). In contrast, artificial neural network represents a more challenging pattern learning problem with inconsistent or unpredictable sampling of datapoints arising when working with smaller datasets (Kouziokas, 2020; Gautam et al., 2020). In the context of water studies, comparative analysis of different machine learning approaches (such as support vector machines, artificial neural network, nearest neighbours, etc.) has shown the robustness and improved performance of support vector machines when compared to other approaches in identifying municipal water network leaks in the presence of different uncertain conditions such as non-stationarity of water use, sensor noise, etc. (Grueiro et al., 2018; Cody, 2020). The major drawback of using a machine learning approach (e.g. support vector machines) is that the trained classifier will only be able to deal with fault alarms that have been experienced by the water distribution system previously. In practical applications, the available data can be limited to a subset of the potential fault alarms that the system has experienced. However, that could be improved by utilizing knowledge of experienced faults or anomalous situations to update the model and provide a more reliable performance monitoring system. As the number of faults or anomalous situations in the fault detection and diagnosis model database increases, the system becomes more valuable and reliable. This study utilizes support vector machines to develop fault detection and diagnosis approaches for non-residential water distribution system for the first time in the literature. To this end, principal component analysis (PCA) and support vector machines (SVM) have not been applied to water distribution systems, the proposed methodology not only to detect and diagnose faults in the water distribution system,

but also involve robust means of identifying outliers within the water time series lead to an advancement in using these computational approaches without explicit knowledge of the system.

Table 1 summarises relevant studies that describe the efficacy and applicability of the principal component analysis and machine learning tools in various engineering domains (municipal water network, nuclear power plants, rotary machines, etc.). The authors did not find any studies which applied principal component analysis and support vector machines methods to building water networks. Previously, literature has focused on stationary systems (and often utilized modelled data). The use of more advanced statistical techniques (e.g. principal component analysis) in conjunction with support vector machines has not been widely explored in water sector. Given that water distribution systems can often exhibit non-stationarity (change in statistical properties over time), the combination of principal component analysis and support vector machines techniques has potential to offer robust applications in water distribution system in non-residential buildings.

187 Table 1: Literature review.

References	Objective	Key points across studies
PCA-based studies		
Horrigan et al., 2018	Statistical fault detection scheme developed by utilizing **commercial building energy performance** data	Proposed statistical-based performance prediction model utilizing univariate data time series. Actual fault scenarios were not studied to validate the fault detection method.
Li et al., 2018	Sensor fault detection in a **nuclear power plant** using principal component analysis method	Fault detection methodology developed based on principal component analysis and detection indices (T^2 and Q - statistics).
Zhang et al., 2017	Assessment of T^2 and Q-statistics for detecting additive and multiplicative faults in multivariate statistical **process monitoring**	Statistical method (principal component analysis) impact on detecting additive and multiplicative faults have been studied in context of fault detection rate. T^2 - statistics performs better than Q - statistics in detecting additive and multiplicative faults.
Ahmed et al., 2012	Fault detection and diagnosis using principal component analysis of **vibration data from reciprocating compressor**	Several diagnostic features extracted, are analysed by PCA to detect leakage faults in a system. The PCA model shows good accuracy in providing detected fault information.
Villegas et al., 2010	Principal component analysis for fault detection and diagnosis, experience with a **pilot plant (two-tanks system**)	PCA methodology developed to identify changes across operating point, assuring model robustness for system non-linearity's
Grueiro et al., 2018	PCA-based leak detection and localization in **water distribution network**	Use of measured data using district metering area (DMA) technique under normal operating conditions. The PCA model shows good accuracy with an average detection rate of over 80%.
Laory et al., 2011 **Posenato et al., 2006**	PCA-based damage detection during continuous static monitoring of civil structures	Use of measured sensor data from linear structures. Two conditions, damage detectability and time to detection, are evaluated with respect to changes in sensor-damage location, dynamic loading, and damage level in structures.
SVM-based studies		
Escofet et al., 2016	Model vs. data based approaches applied to fault diagnosis in potable **water supply networks**	SVM based methodology developed for detecting leakage and to diagnose anomalies in a system. Use of simulated data rather than measured data. Support vector machines shows superiority over artificial neural network in terms of performance, robustness in both normal and noisy conditions. Support vector machines exhibit insensitiveness in detecting leaks in a noisy environment as compared to artificial neural network. Measurement/modelling uncertainties were not studied. Signal decomposition techniques along with support vector machines to extract features for classification. Classification accuracies of discrete wavelet at different levels were calculated and compared. Feature effect on individual fault found missing.
Muralidharan et al., 2014	Fault diagnosis of **monoblock centrifugal pump** using support vector machines	
Nasir et al., 2014	Measurement error sensitivity analysis for detecting and locating **leaks in pipelines** using artificial neural network and support vector machines	
Saberi et al., 2011	Comparing performance and robustness of SVM and artificial neural network for fault diagnosis in a **centrifugal pump**	
Qu et al., 2010	SVM-based **pipeline leakage** detection and pre-warning system	
Mashford et al., 2009	An approach to leak detection in **pipe networks** using analysis of monitored pressure values by support vector machines	
Cody, 2020; **Grueiro et al., 2018**	Compared performance and robustness of different classifiers for detecting leaks in **water distribution networks**	Support vector machines, artificial neural network, nearest neighbours, and Bayes classifiers were used. Use of simulated data rather than measured data. SVM shows advantage over other classifiers in terms of performance, robustness under different uncertain conditions.
Samer et al., 2018	SVM-based single event leak detection in pressurized **water network** pipelines	Use of experimental data at lab-scale setup. Experiments were conducted using one inch and two-inch straight pipelines of materials ductile iron and polyvinyl chloride (PVC). Support vector machines performed with an accuracy of 98% in identifying leaks.
Akil et al., 2019	SVM-based smart monitoring of an institutional building using gas, electricity, and water use data	Use of measured water, electricity, and gas usage data. Support vector machines shows effectiveness in identifying atypical behaviour of an institutional building (such as increased electricity, water use, etc.).
Liu et al., 2019	**Water pipeline** leak detection based on SVM and wireless sensor network	Use of simulated and experimental data to verify the effectiveness of the proposed approach. Integrated conditional probabilities and time-frequency analysis to identify leak signals within the network.
Gautam et al., 2020	SVM-based monitoring and forecasting for leak detection in household **water storage**	Use of measured water use data. Compared actual values against predicted values to identify leaks.

188

This work leverages a case-study building with real-time water consumption data to illustrate the effectiveness of using PCA detection indices (Hotelling T^2-statistics and Q-statistics) in identifying faults in a non-residential building water network. The proposed approach applies the PCA technique to measured time-series water data to (i) identify faults and (ii) classifying the faults into various categories, represented by prediction model output data, to reset and guide operational performance. The key contribution of the proposed methodology is an improvement in the performance monitoring of water distribution system through the identification of faults in early stage for building managers under realistic conditions. Moreover, by utilizing multi-class SVM classification the faults can be classified quickly and effectively with a relatively small amount of training data. This can enable building operators to target predictive maintenance and reduce the occurrence of false alarms, while also developing strategies to reduce unnecessary or inefficient consumption of water.

The proposed fault detection and diagnosis methodology involves analysing water time series and monitoring inevitable changes in a system without the use of geometrical information. In the context of performance monitoring, to differentiate between features of routine and non-routine or anomalous water uses and to limit false detection. Principal component analysis is based on the covariance structure of the data in the PCA space and is sensitive to outlier measurements (Rousseeuw et al., 2018; Hubert et al., 2016). In the presence of outliers, principal components are often attracted towards outlying data points and may not be able to capture the realistic behaviour of routine variation in water distribution system. This can result in unreliable or false fault detection during the process. To address this in the current study, the methodology combines principal component analysis with a distance-distance approach for outlier localization (as discussed in Section 4.1) creating a robust approach where principal components are not influenced by outliers. The proposed methodology is also capable of considering biased uncertainties and non-stationarity of water use which are typical in water distribution system modelling challenges and is in practice seen as state of things rather than process disturbances. Furthermore, in this study the proposed fault detection and diagnosis approach is assessed using real site water distribution system data, with associated outliers and process noise leading to a more robust assessment of the machine learning techniques suitability to detect and diagnose faults in water distribution system.

The objective of the proposed methodology is to obtain a simpler equivalent approach than physics-based models, to reduce the complexities linked with the hydraulic modelling of the non-residential water distribution systems. This proposed approach is designed to be effective in identifying non-routine events and faults of different types (such as equipment malfunctioning, high or low-level imperceptible process faults) not only leakage. Moreover data availability for non-residential building water distribution systems represents a significant and persistent challenge (Matos et al., 2020). In many such cases reliable and objective information about water use and network performance is poor, lacking or otherwise unavailable (Houngbo, 2019). The dataset used in this study is typical of what is available in non-residential buildings, and thus the aim was to develop a fault detection and diagnosis methodology which is capable of operating under these practical data constraints and assessing the effectiveness of a proposed methodology under these practical data constraints.

This paper is arranged as follows: In section 2 materials and methods are presented which include the case study and pumping system details, measurable characteristics, followed by an introduction to principal component analysis, fault detection indices (T^2 and Q-statistics), and fault detection and diagnosis methodology is outlined in Section 3. Later results and discussion are presented in Section 4. Section 5 summarizes the study and provides conclusions and future directions.

2. Methods and materials:

2.1. Case Study:

The case-study site was the Alice Perry Engineering Building at the National University of Ireland, Galway (NUI Galway). The building is located on the west coast of Ireland in an area with a temperate marine climate. The building houses approximately 1,100 students and 100 staff during teaching and exam terms (approximately 26-28 weeks a year in total) and about 150 research, academic, administrative, and technical staff and 50-100 postgraduates during the rest of the year. The building includes lecture halls, classrooms, offices, laboratory facilities, a café, and shower and toilet facilities spread across 14,000 m^2 of floor space on four storeys. Thus, it has a variety of end-uses of its water and hot water systems. The building is managed through a building

management system that collects data on building performance and operational efficiency - including 11 water meters and a number of energy meters. Some of the key water uses include showers and hand wash basins, grey water from a rainwater-harvesting system for toilets and urinals and potable water for the water fountains and the café. This study deals primarily with mains water usage within the building (i.e. hot water system is not considered).

2.2. Building Pumping System:

The mains water system (MWS) in the building is divided into a cold-water system (CWS) and a potable water supply for drinking fountains and for the café. There are two sets of booster pumps (Figure 3), that deliver (i) the cold-water system and (ii) potable water supply. The cold-water system water is then divided into water for grey-water system (GWS) applications and water for other uses such as laboratories, hot water mixing in showers and faucets and non-potable uses in the café. The grey-water system supplies water to toilets and urinal flushing as required by gravity. The building has a large rainwater-harvesting system (RWHS) for supplying grey water. When the rainwater-harvesting system cannot supply sufficient grey water, two holding tanks located on the roof of the building ($8m^3$ capacity each) are topped up by the cold-water system. During the study period, the building rainwater-harvesting system utilised the cold-water system even during periods when rainwater was available due to system faults that remained undetected prior to this study.

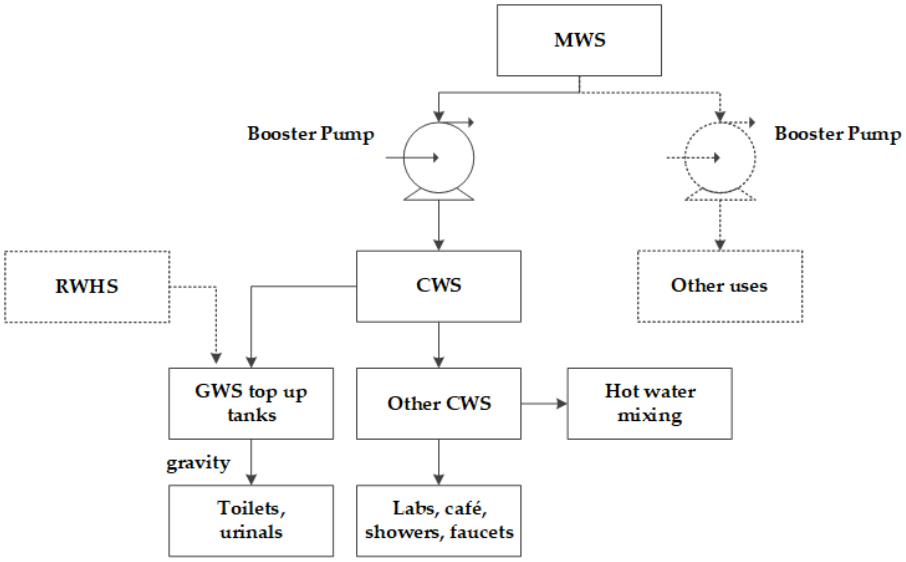

Figure 3: Simplified schematic of water network in the building. The dashed line indicates systems not considered in this study.

2.3. Feature extraction:

To perform fault detection and diagnosis in water distribution systems, measurable characteristics can play a vital role in discriminating between routine and abnormal events in the water network. The characteristics of the system, derived from meter readings, can be used to analyse the water usage baseline and diurnal patterns (or indeed patterns at any chosen time interval for which adequate data is available). For instance, analysing night time flow can assist in detecting continuous flow which can be due to leaks; analysing peak flows can assist in detecting exceptional events. Analysing water consumption at specific daily intervals (working and non-working hours) can help in identifying disruptions to normal routines. The details of the features utilized in this study are listed in Table 2. The daily flow and maximum flow features summarize the water usage over a 24-hour period from midnight to midnight. Whereas the remaining features (working hours, non-working hours, and night time flow) focus on water consumption for particular periods of the day which are linked to how the building is used. In this case, these periods comprise of four working hour time intervals, two non-working hour time intervals and a night time interval. Medium resolution data recorded at 7.5-minute intervals was collected in the case-study building from an inline displacement water meter equipped with a magnetic pulse output and analysed in developing the methodology for fault detection and diagnosis.

Table 2: Features (measured in m^3/hr) used to characterise daily water demand.

	Feature	Description
Total	Daily flow	Average flow in 24 h
	Maximum flow	Highest flow in 24 h
Time of day	Working hours	Average flow between 6 a.m. – 9 a.m.
		Average flow between 9 a.m. – 12 p.m.
		Average flow between 12 p.m. – 3 p.m.
		Average flow between 3 p.m. – 6 p.m.
	Non-working hours	Average flow between 6 p.m. – 9 p.m.
		Average flow between 9 p.m. – 12 a.m.
	Night time flow	Average flow between 12 a.m. – 6 a.m.

For a given year the case-study monitoring period data was divided into Semester 1 - September to December, Semester 2 - January to May and a summer period (June-August). In general, water consumption would be higher in Semesters 1 and 2 than during the summer periods as undergraduate students would not be present during the summer periods. Figure 4a demonstrates the flow pattern at different intervals (Table *2*) during Semester 2 and the summer period under the routine operating condition considered for PCA model training. Limited variation can be observed between working and non-working hour water usage in the summer due to reduced building occupancy as compared to Semester periods. In Figure 4b water consumption patterns are visualized during the summer period. The flow pattern during night time, when the building was not in use, is primarily due to the routine urinal flushing which operate irrespective of the building occupancy. These patterns derived from the water time series, can assist in analysing different faults in the water distribution system.

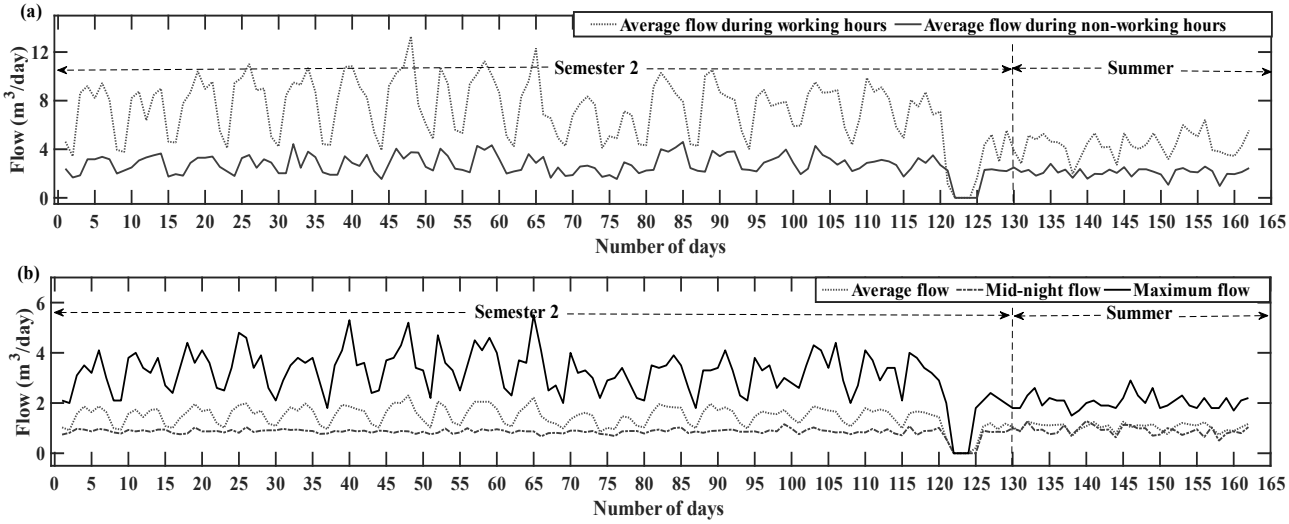

Figure 4: Flow characteristics within the periods of interest (a) Average 7.5 minute interval flow pattern during the 12 working hours and the 6 non-working hours, (b) Maximum, 7.5 minute interval average over 24 hours and night time flow pattern over 6 hours.

This study utilises multivariate statistical approaches (principal component analysis and support vector machines) to attempt to detect patterns in data, to develop a new fault detection and diagnosis approach for non-residential building water distribution systems. The theory behind the statistical methods used are outlined in the next section. It is noted that the primary advantages of the combined principal component analysis and support vector machines approaches are the ability to compress large data sets without losing data definition. In this case it can enable analysis of faults across a number of water usage features, both of which reduce the likelihood of false positives and false negatives within the fault detection and diagnosis system. In addition, support vector machine facilitates classification of faults, providing a greater level of information to building managers.

3. Theory:

3.1. Principal Component Analysis:

Computation of all feature values for water consumption, gives a data matrix of $X \in \mathrm{R}^{n\times m}$ whereby *n* rows represent the values for all features (Table 2) within a single day and *m* columns represent a single feature (such as maximum flow, average flow, etc.) over the observation with mean zero and unit variance.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix} = (w_1 \quad w_2 \quad \cdots \quad w_j \ldots \quad w_m) \tag{1}$$

where row vector X_i represents all feature measurements (i.e. daily flow, etc.) at specific time intervals and column vector w_j represents single feature measurements (average flow, night time flow, etc.) in a series of days. The data matrix $X \in \mathrm{R}^{n\times m}$ is then transformed by utilizing singular value decomposition as per (Jackson & Mudholkar, 1979; Zenobi et al., 2011; Sheriff et al., 2017).

$$X = TP^T + E \tag{2}$$

where $T = [t_1, t_2, t_3, \ldots t_m] \in \mathrm{R}^{n\times m}$ is a matrix of transformed variables, where each column represents the score vectors and the i^{th} eigenvalue equals the square of the i^{th} singular value (i.e. $\lambda_i = \sigma_i^2$). $P = [p_1, p_2, p_3, \ldots p_m] \in \mathrm{R}^{m\times m}$ is the matrix of orthogonal vectors, where each column is populated by the eigenvectors associated with the covariance matrix of the data matrix $X \in \mathrm{R}^{n\times m}$. *E* is the residual matrix which ideally contain noise in the data. Typically, most of the data variance is contained in the principal components with larger eigenvalues, while the remaining principal components are considered as measurement noise which can be removed by reducing the data dimension (Grueiro et al., 2018; Laory et al., 2011). The covariance matrix *S* quantifies the amount of linear correlation between all possible combinations of features within the dataset and can be computed as (Sengupta & Kundu, 2016; Rosen, 2001; Jackson & Mudholkar, 1979):

$$S = \frac{1}{n-1} X^T X = P \Lambda P^T \tag{3}$$

where Λ is the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3 \ldots \lambda_m) \in \mathrm{R}^{m\times m}$ containing non-negative eigenvalues related to *m* principal components and magnitudes in a descending order $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_m \geq 0$. The model built by principal component analysis contains the same number of principal components as the original number of features in the data matrix $X \in \mathrm{R}^{n\times m}$. In the case of water distribution systems, it may contain features (such as average flow, average flow during working hours, average flow during non-working hours, etc.) that are highly correlated. In this study, the number of components is selected based on the cumulative percent variance (CPV) due to its computational simplicity and wide adoptability (Sheriff et al., 2017). It provides a good estimate of the number of principal components that need to be retained for most practical applications. Cumulative percent variance is a measure of the percentage variance CPV(*r*)≥ 90% captured by the first *r* principal components (Zumoffen & Basualdo, 2008; Johnson & Wichern, 2007):

$$CPV(r) = \frac{\sum_{i=1}^{r} \lambda_i}{trace(S)} 100 \tag{4}$$

3.1.1. Score and orthogonal distance:

To analyse the linear combination of water time series that restrain valuable information or outliers; the score distance SD_i and the orthogonal distance OD_i of each data point to the PCA subspace are given by (Hubert et al., 2005; Harris et al., 2014; Rousseeuw et al., 2018).

$$SD_i = \sqrt{\sum_{j=1}^{r} \frac{t_{ij}^2}{\lambda_j}} \quad (5)$$

330

$$OD_i = \left\| x_i - \hat{\mu} - P_{m,r} t_i' \right\| \quad (6)$$

331 where t_{ij} is the score value of each data point, λ_j is the non-negative eigenvalues and $P_{m,r}$ is the eigenvectors matrix
332 related to *r* principal components, $\hat{\mu}$ is the mean of the covariance matrix. Assuming the data follows a multivariate
333 normal distribution. The cut-off for the data points were estimated using Chi-square distribution $\sqrt{X^2_{r,0.975}}$. Where,
334 $X^2_{r,0.975}$ is the 97.5% quantile of the Chi-squared distribution with *r* principal components (Rousseeuw & Hubert,
335 2018).

336 *3.2. Fault detection indices:*

337 *3.2.1. Hotelling T^2-statistics:*

338 T^2-statistics represents the major variation in the data (Garcia-Alvarez, 2014). The T^2-statistics of the i^{th} sample
339 or observation *x* can be expressed by (Mujica et al., 2011; Qin, 2003) as,

$$T_i^2 = x_i^T P \, \Lambda_r^{-1} P^T x_i \quad (7)$$

340 where, Λ_r^{-1} is the diagonal matrix containing the eigenvalues related to retained principal components, x_i is the
341 data vector of the i^{th} observation and *P* contains the loading vector associated with the *r* columns. Under routine
342 process conditions, the data follow a multivariate normal distribution, the T^2-statistics is related to an *F*-
343 distribution considering that the population mean, and covariance are estimated from data (Qin, 2003).

344 The control limit T_α^2 can be obtained by *F*-distribution as follows.

$$T_\alpha^2 = \frac{r(n-1)}{n-r} F_{\alpha(r,n-r)} \quad (8)$$

345 where *n* is the number of observations in the data, *r* is the number of retained principal components, $F_{\alpha\,(r,\,n-r)}$ is the
346 *F*-distribution. In the current study, the α-control limits are calculated at a confidence level of 95%. The T^2-
347 statistics can be interpreted as measuring the systematic variations of the process. Violation of routine conditions
348 would indicate that the systematic variations are outside normal operating bounds, and thus the data set may
349 indicate a fault or non-routine process condition (Vieira et al., 2014). The new observation is considered to be
350 normal if it satisfies the following condition.

$$\begin{cases} T_i^2 < T_\alpha^2 - - - Normal \\ T_i^2 \geq T_\alpha^2 - - - System\ alarm \end{cases} \quad (9)$$

351 *3.2.2. Q-statistics or squared prediction error:*
352 The Q-statistics also known as squared prediction errors, measure the variability of the observation that violates
353 the routine process correlation (with small or moderated magnitudes) not accounted by the principal component
354 subspace. In case of water distribution systems, continuous flow usually does not impact on the overall water
355 consumption in a short period and so may not be detected by conventional methods, however a low-grade fault,
356 such as minor continuous flows, can result in undesirable water loss and accumulate to significant losses. The
357 portion of the measurement space corresponding to the lowest (*m* - *r*) eigenvalues are thus monitored (Ballabio,
358 2015; Qin, 2012).

$$Q_i = x_i^T (I - P_i P_i^T) x_i \quad (10)$$

359 where *I* is the identity matrix. The control limit Q_α can be computed from its approximate distribution (Ballabio,
360 2015; Qin, 2012).

$$Q_\alpha = \theta_1 \left[\frac{h_0 c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 \ \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \quad (51)$$

$$\theta_i = \sum_{j=r+1}^{n} \lambda_j^{2i} \quad (62)$$

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2} \quad (73)$$

where r is retained principal components inside the model and c_α is the standard normal deviation with the upper (1 - α) percentile. When an unusual event occurs, and it produces a change in the covariance structure of the model, it will be detected by a high value of Q. The new observation is considered to be normal if it satisfies the following condition.

$$\begin{cases} Q_i < Q_\alpha --- Normal \\ Q_i \geq Q_\alpha --- \text{System alarm} \end{cases} \quad (84)$$

When the Q_i of the new experimental trial violates the Q_α control limit, this is indicative of a fault or a non-routine event. The value of Q_α is defined with an assumption that the observation data is multivariate normally distributed and time-independent (Harrou et al., 2013). The value Q-statistics are small and consequently more sensitive than the T^2-statistics. This characteristic makes the Q-statistics ideally suited to detecting minor variation within the system behaviour. It evaluates the variation of the new incoming water data which are not accounted for by the principal component subspace. On the contrary, the T^2-statistics require significant variation in the system behaviour to be measurable (Mujica et al., 2011).

3.2.3. Fault alarm reduction:

In order to isolate actual fault alarms from system alarms, and subsequently reduce false alarms in the fault detection and diagnosis system, the time series nature of the data over periods of two days or more were considered. Non-residential public use buildings such as the one explored herein, are occasionally used for large events which result in significant increases in water consumption (i.e. conferences, seminars, etc. on a given day). Thus, fault alarms in the proposed framework are raised only when system alarms persist over two or more consecutive days. Thus, when the analysed data exceeded the α-control limits (Eq. 15) for more than two days a fault alarm is triggered. The two-day time period was selected based on a historical analysis of the events within the building (conferences, seminars, workshops, etc.). The nature of the time series alarm system could vary for different buildings based on water consumption patterns and event patterns and can be established from a high-level water audit. It is also noted that this framework step can be removed for buildings where occasional large occupancy events do not occur (i.e. food processing facilities, manufacturing facilities, etc.).

$$\begin{cases} t_{acceptable} < t_{limit} --- Normal \\ t_{acceptable} \geq t_{limit} --- Indicative\ of\ fault \end{cases} \quad (95)$$

3.3. Fault diagnosis:

Support vector machine (SVM) is a supervised learning technique widely used for classification and regression. SVM creates a decision boundary in between classes by mapping the training data (through kernel functions) onto a higher dimensional space, and then obtaining the maximum margin hyperplane within that space. Finally, this hyperplane can be regarded as a classifier. To achieve multi-class classification, error-correcting output code (ECOC) was used. ECOC represents a powerful framework to deal with multi-class classification problems based on combining binary classifiers (such as support vector machine, neural networks, decision trees, etc.) (Bagheri et al., 2012; Lin, 2018). A detailed explanation of the support vector machine technique can be found in (Bishop, 2013). The support vector machine technique has been used herein due to its wide use in different engineering sectors (such as thermal power plants, centrifugal pumps, bearing, etc.) in context of classification (Sabri et al., 2017; Xiao, 2016; Bayar et al., 2015; Chen et al., 2011, Hmeidi et al., 2008).

Fault diagnosis of the system alarms in this study was carried out based on the domain knowledge and process historical information of faults within the building. In this study, SVM was trained by considering the labelled flow water data. Samples with different system alarm labels (outlined in Table 3) were obtained from the training data. It is noted that fault data across various faults was somewhat limited in the training data due to lack of occurrence of these faults over the monitoring period.

Moreover, prior to detecting a routine or faulty condition, cross-validation was carried out to optimize classifier's hyper-parameters and asses the performance of the classifier. It is noted that SVM is less effective when the data is noisy and contains overlapping data points (Lin, 2018). The goal of cross-validation is to gauge the generalizability of an algorithm and to prevent overfitting by minimizing the influence of noisy data (Mutasa et al., 2020; Gupta, 2019; Mudry & Tjellström, 2011). In this study, the SVM classifier was trained by conducting stratified 10-fold cross validation to minimize the generalization error (minimizing the error associated with the biasness and variance of data). It also tends to provide less biased estimation of the accuracy. In this case a classification accuracy of 80.95% was achieved during training. The summary of the fault detection and diagnostic approach is outlined in Figure 5.

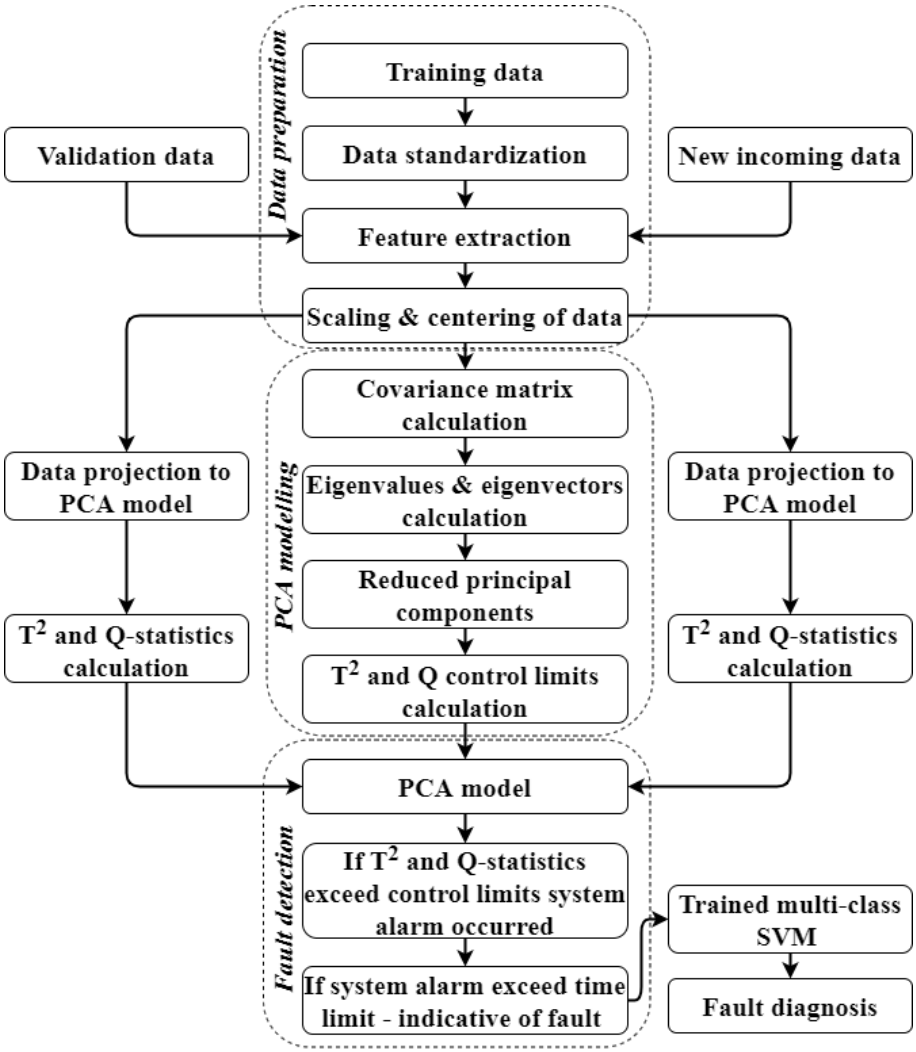

Figure 5: Summary of the proposed fault detection and diagnosis approach.

4. Results and discussion:

4.1. Training:

The water usage data was aggregated into hourly flow traces (i.e. 8 readings at 7.5-minute intervals averaged over one hour) to analyse the usage characteristics at the pilot site. In a number of cases, data points had overlapping timestamps due to metering errors and these were removed from the training dataset. Real water time series data

often contains values that exhibit atypical usage characteristics and do not follow expected or routine patterns. Such data points are normally identified as outliers (Harris et al., 2014). These outliers can be due to measurement error, variation in the water use, and process noise, but could assist in stabilizing the sensitivity of the PCA model. Thus, some outliers in the dataset may contain valuable information which could be lost if all outliers are removed, while others (i.e. measurement error and process noise) can negatively affect the model accuracy and could be removed. To localize outliers in this context, the dataset was analysed by plotting distances (i.e. score and orthogonal distances) from the centroid of the covariance data structure (Figure 6). The score distance was measured within a PCA space, while the orthogonal distance was measured between data each point and its projection in the *r*-dimensional space (Rousseeuw et al., 2018).

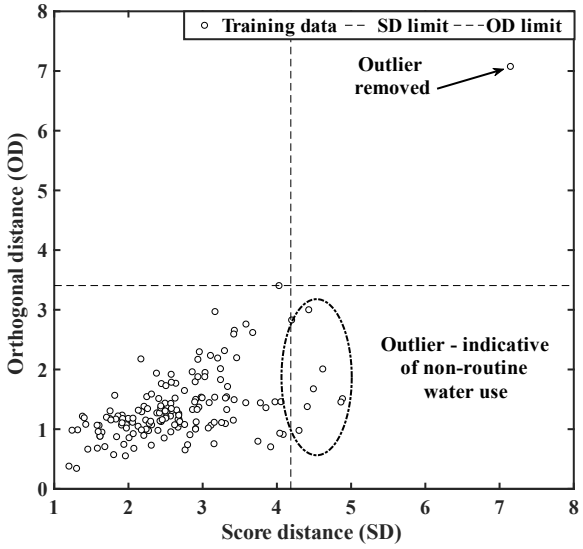

Figure 6: Distance-distance plot for the training dataset.

The PCA model and the SVM model were trained using six months of data from January to June (which incorporated a Semester of teaching and exams); the calculated PCA monitoring statistics (T^2-statistics and Q-statistics) are shown in Figure 7 (dashed lines). As can be seen from Figure 7, a number of points lie above the α-control limits, constituting system alarms. These are due to a combination of non-routine events in the building (e.g. event at 122 days) and routine statistical variation related to the inherited definition of the α-control limit, which is linked to false alarm probability (Li et al., 2019). Importantly, none of these system alarms in the training data would trigger a fault alarm in the fault detection and diagnosis framework as none of the system alarms occur on consecutive days (as discussed in Section 3.2.3). It is also noted that metering fault data and one obvious fault data point were removed from the training data set, in line with standard threshold development procedure.

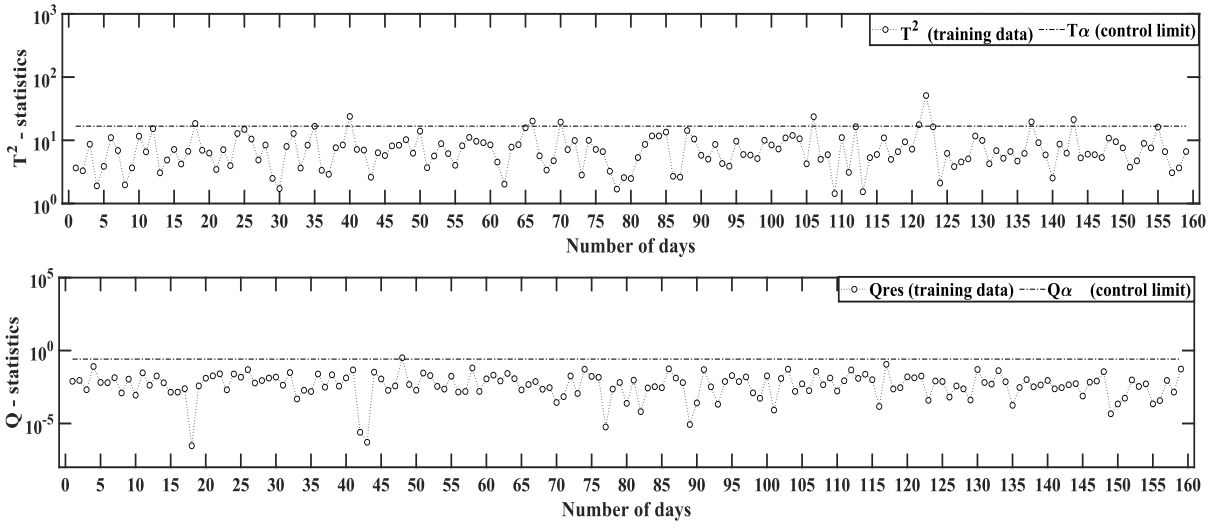

Figure 7: The time evolution of T^2-statistics and Q-statistics on a semi-logarithmic scale for the training data (PCA model).

4.2. Validation:

Having trained the PCA model, a model validation was conducted on academic Semester 1 water consumption data (September-December) to assess the model's ability to detect faults. Importantly, it was known that this Semester 1 data contained an actual fault in the building water distribution system. A previous study (Clifford et al., 2018) focused on the use of low-resolution meter data to detect various flow signatures associated with different end-use cases in this building. This study demonstrated faults had occurred in the rainwater-harvesting system, which was eventually found to be due to a faulty valve on the inlet to one of the storage tanks on the roof top of the building. Further inspection of meter data from the period concerned indicated that the fault caused excess metered water consumption of 3.5 m^3/day in the building (as the cold water system supply was engaged to fill the rainwater top-up tanks even though rainwater was available). Figure 8a illustrates the total daily usage of water over the training period and during the Semester 1 period (validation period) where the fault occurred. Figure 8b & c show the PCA model output for the validation period, which utilised the Hotelling T^2-statistics and Q-statistics α-control limits calculated during the training phase.

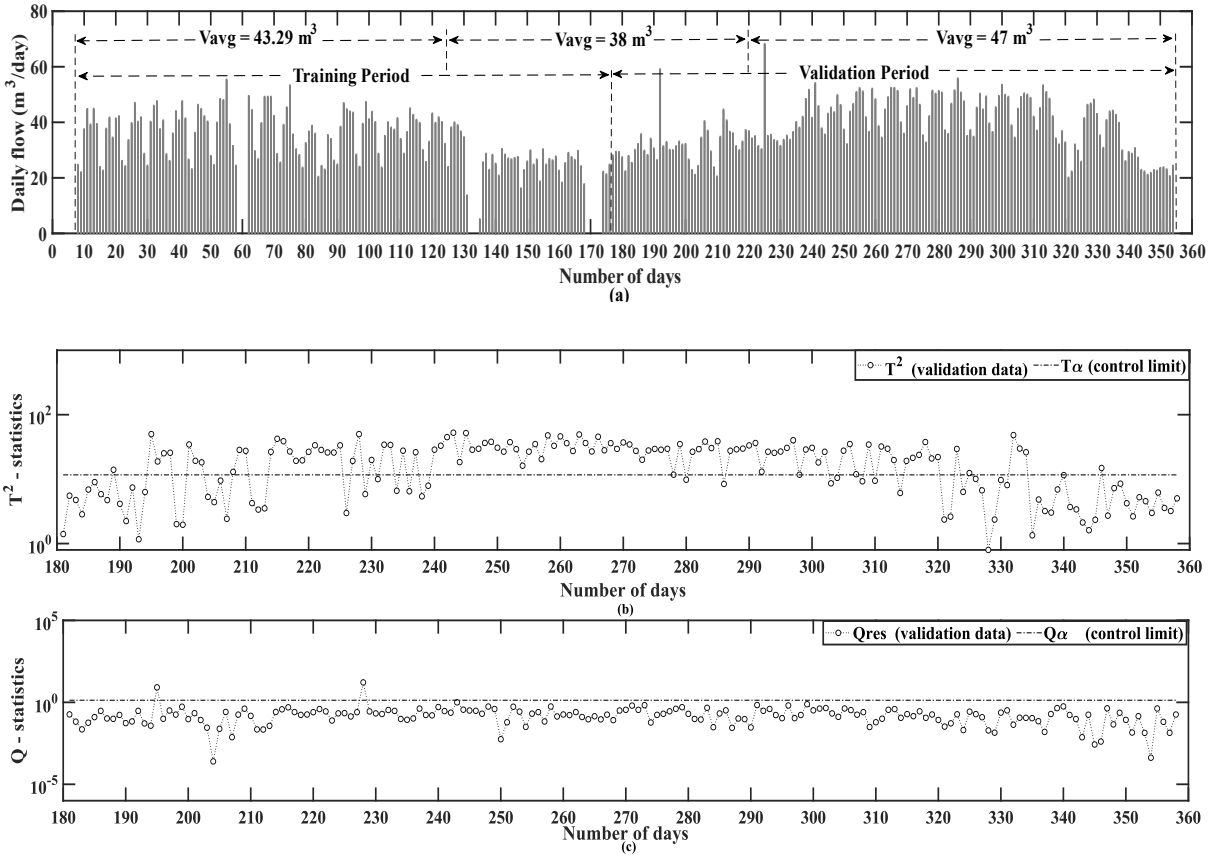

Figure 8: (a) Daily usage of total water use in the building, (b & c) The time evolution of T^2-statistics and Q-statistics on a semi-logarithmic scale for the testing data (PCA model validation).

As can be seen from the plots, the T^2-statistics for the Semester 1 data breach the α-control limit T^2_α for a large portion of the monitoring period, resulting in a significant number of system alarms. As many of these system alarms occurred on consecutive days, fault alarms were triggered in the PCA fault detection and diagnosis framework in accordance with the false alarm reduction step (Section 3.2.3). While the T^2-statistics were effective in detecting the water distribution system faults (Figure 8b), the Q-statistic did not pick up the fault in question with few Q values breeching Q_α. This is due to the fact that the T^2-statistics is more suited to detecting faults relating to higher flows, while the Q-statistics is targeted towards detecting lower flow reading faults. To gain further insight into the PCA model performance, the results were compared to those obtained using the univariate

method in (Clifford et al., 2018). The PCA model outperformed the univariate model over the validation period, detecting 25% more system alarms and triggering four more fault alarms. When these additional system alarms were checked back against the historical records for the building, they correctly corresponded to known faults or known non-routine events. Comparisons between the performance of the univariate fault detection and diagnosis approach and the PCA and SVM fault detection and diagnosis approach proposed herein is explored in more detail in Section 4.3 of this paper.

4.3. Evaluation:

Having trained and validated the fault detection model, its performance was evaluated by considering data for the year following the validation period. As can be seen in Figure 9, a large number of data points (272 data points) exceeded the α-control limit and were thus labelled as system alarms i.e. non-routine events or due to faults in the system. Of these 272 system alarms, sixteen sets of fault alarms were raised (dashed line boxes in Figure 9). As per the Section 3.2.3 these fault alarms occur when system alarms (caused by values above the α-control limit) persisted for two consecutive days or more. Several non-routine events were also observed over the 1-year period (i.e. whereby there was a single 1 day above α-control limit). These high usage peaks raised system alarms but did not trigger a fault alarm in this fault detection and diagnosis approach.

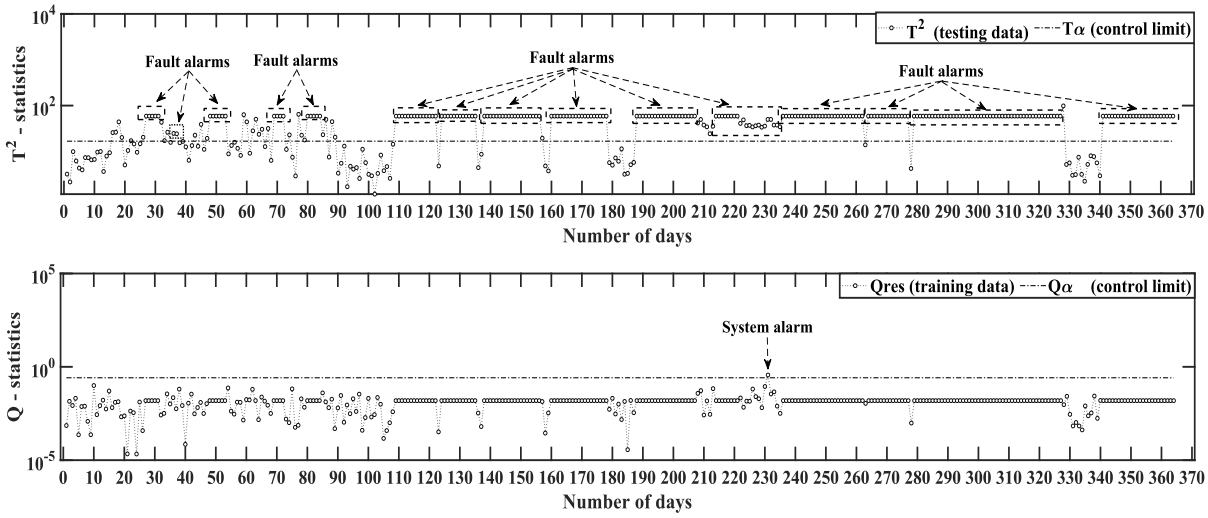

Figure 9: The time evolution of T^2-statistics and Q-statistics on a semi-logarithmic scale for the new testing data (PCA fault detection).

Having detected the system alarms using the PCA approach, the model then used SVM to classify each system alarm. The SVM classifier checked system alarms to ensure they were not in fact routine flow events i.e. some events which are close to the α-control limit may be classified as routine flow. The remaining system alarms were classified as 1) non-routine events (non-consecutive system alarms), 2 metering error, or 3) excess usage indicatives of fault. This is shown in Table 3. It is noted that, in this case, the SVM could not be trained to classify continuous flow events for this case study as no such events occurred during the training data period.

Table 3: System alarm classification labels.

Labels	Description	
1	Non-routine event	Unusual water uses in working or non-working hours
2	Metering error	Error due to equipment malfunctioning / zero water use during the time period
3	Excess usage	Exceptional water uses during the time period indicative of water distribution system fault

Importantly, the SVM classification identified the majority of the sixteen fault alarm sets (sequential α-control limit breeches), fifteen were due to metering errors (the previous fault identified with the rainwater-harvesting system was explored as a possible issue but was found to have been resolved). On further investigation the "metering error" was in fact found to be the result of flow data not being logged and thus there was no data for these periods (which in this approach are equivalent to "zero" flow days). For days with system alarms these were attributed to workshops, infrequent laboratory events, conferences, etc. that occurred on the given days. In retrospect it was not possible to identify the exact nature of the error.

These results are reflective of situations where such data sets can have a significant portion of missing data for a range of reasons including: malfunctioning of monitoring equipment (the replacement of which is seldom a short-term investment priority), lapses in monthly or annual contracts between building owner and third party data monitoring companies, etc. It is noted that if the proposed fault detection and diagnosis system was in place over the evaluation data period the faulty meter would have been identified and subsequently replaced.

Figure 10 represents the performance of the classifier across the three system alarm classifications via a confusion matrix. Confusion matrices (aka contingency tables) reveal how the classifier mislabels (or confuses) system alarms and can be used to summarise the performance of a classifier with respect to test data (Vitter & Webber, 2018). In this matrix, the main diagonal elements represent how often a certain system alarm label is classified correctly (green), while the other boxes show the classification results for misclassification. It is noted that the correct classification of the data for this comparison was determined from time consuming manual interrogation of the historical test data for all system alarms over the monitoring period. In general, Figure 10 shows that the classifier performs well given the relatively short training dataset of six months. The extent of error in classification is a function of a) the difficulty in identification of a particular class, and b) the small sample sizes of the different system alarms present in the training data. For instance, the training data contained only three metering errors. The classifier can be easily trained for this error however, as metering error readings nearly always show zero flow or a series of days with the same flow (which is generally highly unlikely). This results in 100% correct classification for metering error as shown in Figure 10. The training data also contained three non-routine events; however, this data set was only sufficient to ensure 50% classification accuracy for non-routine events. This is due to the fact that many non-routine events are borderline flow events, which are close to the T^2 and Q α-control limits and are thus highly susceptible to misclassification. For excess usage, which is perhaps the most important classification as it is indicative of a fault in the water distribution system, the classifier was 67% accurate. This success rate was obtained with only two excess usage data events in the training data. This is partially due to the fact that excess usage events tend to have easily distinguishable characteristics such as high flows and are thus easier to distinguish from routine operation than say non-routine events. While overall classifier performance was good, the analysis clearly indicated that a greater number of system alarms in the training data would facilitate improved classification going forward. Thus, from a practical viewpoint, a classifier can be continually improved during the performance monitoring period, through continued updating of the system alarm training dataset.

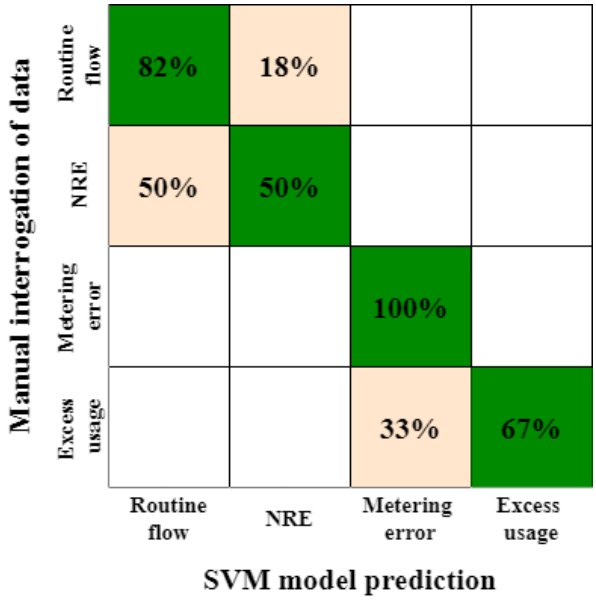

Figure 10: Confusion matrix of an actual and predicted system alarm labels.

4.4. Effectiveness of fault detection and diagnosis:

In order to gain further insight into the performance of the fault detection and diagnosis approach proposed herein, the results over the evaluation period were compared to those obtained from the more basic univariate approach for the same data set. The findings from this comparison are presented graphically in Figure 11a. As discussed, 272 system alarms were detected by the proposed fault detection and diagnosis approach. As shown in Figure 11, 248 of these mapped directly onto the univariate model results, meaning the PCA approach detected an extra 24 system alarms over the univariate method. Manual interrogation of the historical data revealed that of these 24 additional system alarms, 13 corresponded to non-routine events, while 11 constituted false system alarms (i.e. examination of the historical data indicated the water distribution system in the building was operating correctly at the time of the system alarm). As shown in Figure 11b however, the performance of the proposed approach was further enhanced by the SVM component of the model. The SVM classification resulted in 11 of the false system alarms being re-classified as routine flow. Consequently, overall, the method proposed herein identified all the univariate model system alarms, detected 13 additional system alarms successfully, and revealed 3 were found to be false alarms. The proposed approach also facilitated accurate classification of the system alarms (over 90% accuracy), providing a greater level of information to the building manager. Again it is noted that continuous flow alarms could not be classified due to a lack of these faults in the training data, and while the proposed approach has been shown to have advantages over the standard univariate approach, its performance would be even further enhanced through provision of larger training data sets.

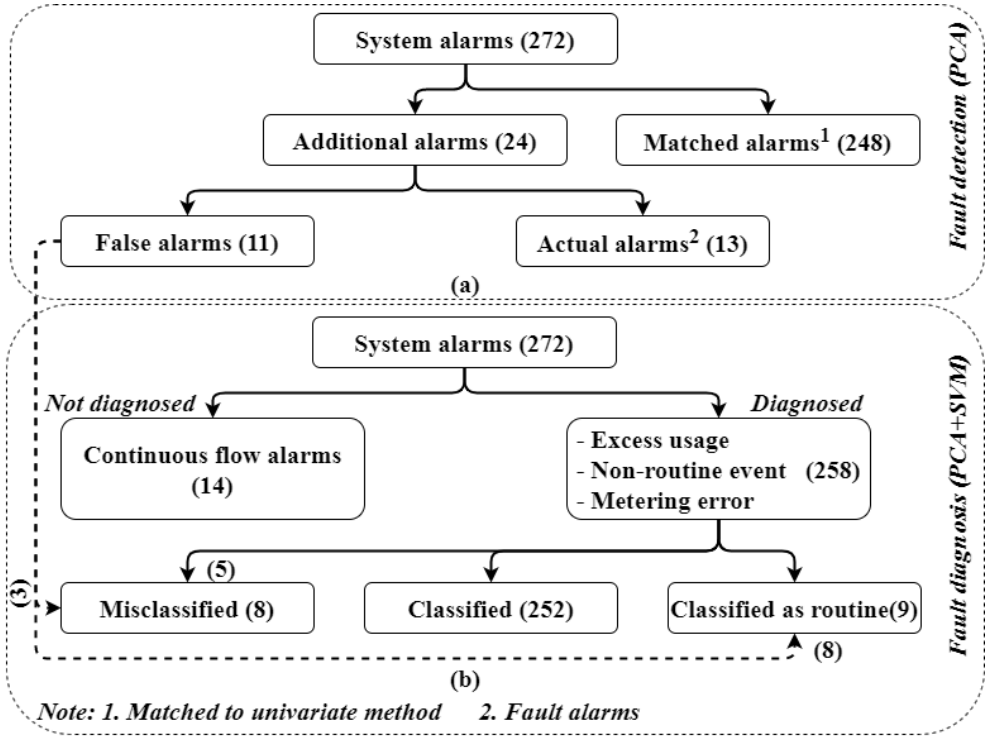

Figure 11: Description of (a) system alarm detection results and (b) system alarm diagnosis results.

5. Conclusion and future directions:

Principal component analysis together with a multi-class classifier support vector machine were found to be useful in detecting and diagnosing faults in building water networks. Classical and advanced approaches have resolved many fault detection and diagnosis problems in water networks, but when it comes to building level multivariate systems or special cases (for instance substantial variation in the water consumption data), these approaches are not entirely effective. This study investigated a proposed fault detection and diagnosis approach for non-residential building water distribution system which combines the detection indices (T^2 and Q-statistics) and multi-class support vector machine. Hotelling T^2-statistics and Q-statistics were used to detect system alarms in the incoming data and the latter multi-class support vector machine along with error-correcting output code was trained for system alarm classification. The results demonstrated promising capabilities of the proposed fault detection and diagnosis approach. When compared to the standard univariate approach, a greater number of system alarms were detected and found to have occurred. The multi-class support vector machine also allowed these system alarms to be classified, providing a greater level of information to building managers, which may avoid unnecessary emergency shutdown in industrial applications. Thus, the comparative study has shown that the proposed approach performs better than standard univariate approach. While the proposed approach has good capability in detecting and diagnose faults in complex non-residential water distribution systems, it also requires less computational resources compared to univariate approach.

Another important characteristic is adaptability. Once a new fault or anomalous situation is identified, the model can be updated to incorporate this additional information. Adaptation will make system more reliable and valuable in time and allow for detection and diagnosis of further anomalies in the water distribution systems. The approach would be further enhanced by extending training data periods beyond the six-month used in this study, especially for system alarm classification. The authors do not envisage that this would be a problem in practice.

Future work will aim to improve some practical aspects of the model, such as reducing false alarm probability and improving system alarm classification through enhanced training datasets. The latter is likely to be more challenging as data from known faults can be difficult to obtain as there may be limited faults in the training data or many historical faults may be of a similar classification. Furthermore, introducing faults to functional systems for the purposes of data collection is somewhat impractical. It would also be interesting and potentially very useful

to develop a hybrid performance monitoring system (i.e. combining model-based and data-driven approaches), which could combine the strengths of each approach.

Acknowledgements

This paper has emanated from research conducted as a part of Energy Systems Integration Partnership Programme (ESIPP) project with the financial support of Science Foundation Ireland under the SFI Strategic Partnership Programme Grant Number SFI/15/SPP/E3125.

References:

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Abdulshaheed, A., Mustapha, F., & Ghavamian, A. (2017). A pressure-based method for monitoring leaks in a pipe distribution system: A Review. *Renewable and Sustainable Energy Reviews*, *69*(May 2015), 902–911. https://doi.org/10.1016/j.rser.2016.08.024

Ahmed, M., Baqqar, M., Gu, F., & Ball, A. D. (2012). Fault detection and diagnosis using Principal Component Analysis of vibration data from a reciprocating compressor. *Proceedings of the 2012 UKACC International Conference on Control, CONTROL 2012*, (September 2012), 461–466. https://doi.org/10.1109/CONTROL.2012.6334674

Ait-Izem, T., Harkat, M. F., Djeghaba, M., & Kratz, F. (2018). On the application of interval PCA to process monitoring: A robust strategy for sensor FDI with new efficient control statistics. *Journal of Process Control*, *63*, 29–46. https://doi.org/10.1016/j.jprocont.2018.01.006

Akil, M., Tittelein, P., Defer, D., & Suard, F. (2019). Statistical indicator for the detection of anomalies in gas, electricity and water consumption: Application of smart monitoring for educational buildings. *Energy and Buildings*, *199*, 512–522. https://doi.org/10.1016/j.enbuild.2019.07.025

Alsaydalani, M. O. A. (2017). Influence of cavitation phenomenon on the hydraulic behavior of leaks in water distribution systems. *Journal of Water Supply: Research and Technology - Aqua*, *66*(5), 327 LP – 339. Retrieved from http://aqua.iwaponline.com/content/66/5/327.abstract

Bagheri, M. A., Montazer, G. A., & Escalera, S. (2012). Error correcting output codes for multiclass classification: Application to two image vision problems. *AISP 2012 - 16th CSI International Symposium on Artificial Intelligence and Signal Processing*, (Aisp), 508–513. https://doi.org/10.1109/AISP.2012.6313800

Balaras, C., Dascalaki, E., Droutsa, K., Micha, M., Kontyiannidis, S., & Argiriou, A. (2017). Energy use Intensities for Non-Residential Buildings. *Proceedings of the 48th International HVAC&R Congress*, (December), 369–389. https://doi.org/10.24094/kghk.017.48.1.369

Ballabio, D. (2015). A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. *Chemometrics and Intelligent Laboratory Systems*, *149*(November), 1–9. https://doi.org/10.1016/j.chemolab.2015.10.003

Bayar, N., Darmoul, S., Hajri-Gabouj, S., & Pierreval, H. (2015). Fault detection, diagnosis and recovery using Artificial Immune Systems: A review. *Engineering Applications of Artificial Intelligence*, *46*(November 2017), 43–57. https://doi.org/10.1016/j.engappai.2015.08.006

Benaicha, A., Mourot, G., Benothman, K., & Ragot, J. (2013). Determination of principal component analysis models for sensor fault detection and isolation. *International Journal of Control, Automation and Systems*, *11*(2), 296–305. https://doi.org/10.1007/s12555-012-0142-x

Bishop, C. M. (2013). *Pattern Recognition and Machine Learning*. *Journal of Chemical Information and Modeling* (Vol. 53). https://doi.org/10.1117/1.2819119

Bruton, K., Raftery, P., Kennedy, B., Keane, M. M., & O'Sullivan, D. T. J. (2014). Review of automated fault detection and diagnostic tools in air handling units. *Energy Efficiency*, *7*(2), 335–351. https://doi.org/10.1007/s12053-013-9238-2

Burak Gunay, H., Shen, W., & Newsham, G. (2019). Data analytics to improve building performance: A critical review. *Automation in Construction*, *97*(June 2018), 96–109. https://doi.org/10.1016/j.autcon.2018.10.020

Chen, K. Y., Chen, L. S., Chen, M. C., & Lee, C. L. (2011). Using SVM based method for equipment fault detection in a thermal power plant. *Computers in Industry*, *62*(1), 42–50. https://doi.org/10.1016/j.compind.2010.05.013

Chen, T. (2010). On reducing false alarms in multivariate statistical process control. *Chemical Engineering Research and Design*, *88*(4), 430–436. https://doi.org/10.1016/j.cherd.2009.09.003

Clifford, E., Mulligan, S., Comer, J., & Hannon, L. (2018). Flow-Signature Analysis of Water Consumption in

Nonresidential Building Water Networks Using High-Resolution and Medium-Resolution Smart Meter Data: Two Case Studies. *Water Resources Research*, *54*(1), 88–106. https://doi.org/10.1002/2017WR020639

Cody, R. (2020). Acoustic Monitoring for Leaks in Water Distribution Networks by.

Cody, R. A., & Narasimhan, S. (2020). A field implementation of linear prediction for leak-monitoring in water distribution networks. *Advanced Engineering Informatics*, *45*(April). https://doi.org/10.1016/j.aei.2020.101103

Connor, B. O., & Murphy, C. (2017). Irish Water.

Cuguero Escofet, M. À., Quevedo, J., Alippi, C., Roveri, M., Puig, V., & García, D. (2016). Model- vs. data-based approaches applied to fault diagnosis in potable water supply networks. *Journal of Hydroinformatics*, (May), 1–20. https://doi.org/10.2166/hydro.2016.218

D'Agostino, D., Cuniberti, B., & Bertoldi, P. (2017). Energy consumption and efficiency technology measures in European non-residential buildings. *Energy and Buildings*, *153*(2017), 72–86. https://doi.org/10.1016/j.enbuild.2017.07.062

Danacova, M., Fencik, R., & Nosko, R. (2016). Historical Development of the Permanent Gully Erosion - Case Study Tura Luka. *Water, Resources, Forest, Marine and Ocean Ecosystems Conference Proceedings, Vol I*, (June), 391–398. https://doi.org/10.5593/sgem2016B31

Datta, S., & Sarkar, S. (2016). A review on different pipeline fault detection methods. *Journal of Loss Prevention in the Process Industries*, *41*, 97–106. https://doi.org/10.1016/j.jlp.2016.03.010

EPA. (n.d.). Water use efficiency in buildings. In *EPA of USA* (pp. 48–71). https://doi.org/10.1002/9781118456613

European Commission - DG Environment. (2012). Water Performance of Buildings, (August), 154.

Garc\ia-Alvarez, D. (2014). Fault detection using Principal Component Analysis (PCA) in a Wastewater Treatment Plant (WWTP), (January). Retrieved from http://sntk09en.guap.ru/sntk09en/main/docs/Alvarezisa.pdf

Gautam, J., Chakrabarti, A., Agarwal, S., Singh, A., Gupta, S., & Singh, J. (2020). Monitoring and forecasting water consumption and detecting leakage using an IoT system. *Water Supply*, *20*(3), 1103–1113. https://doi.org/10.2166/ws.2020.035

Geng, Y., Ji, W., Wang, Z., Lin, B., & Zhu, Y. (2018). A review of operating performance in green buildings: Energy use, indoor environmental quality and occupant satisfaction. *Energy and Buildings*, *183*, 500–514. https://doi.org/10.1016/j.enbuild.2018.11.017

Gharsellaoui, S., Mansouri, M., Trabelsi, M., Refaat, S. S., & Messaoud, H. (2020). Fault diagnosis of heating systems using multivariate feature extraction based machine learning classifiers. *Journal of Building Engineering*, *30*(September 2019), 101221. https://doi.org/10.1016/j.jobe.2020.101221

Gupta, S. (2019). 2019_Dealing with Noise Problem in Machine Learning Data-sets.pdf.

Harris, P., Brunsdon, C., Charlton, M., Juggins, S., & Clarke, A. (2014). Multivariate Spatial Outlier Detection Using Robust Geographically Weighted Methods. *Mathematical Geosciences*, *46*(1), 1–31. https://doi.org/10.1007/s11004-013-9491-0

Harrou, F., Nounou, M. N., Nounou, H. N., & Madakyaru, M. (2013). Statistical fault detection using PCA-based GLR hypothesis testing. *Journal of Loss Prevention in the Process Industries*, *26*(1), 129–139. https://doi.org/10.1016/j.jlp.2012.10.003

Hmeidi, I., Hawashin, B., & El-Qawasmeh, E. (2008). Performance of KNN and SVM classifiers on full word Arabic articles. *Advanced Engineering Informatics*, *22*(1), 106–111. https://doi.org/10.1016/j.aei.2007.12.001

Houngbo, G. F. (2019). *The United Nations world development report 2019. Leaving no one behind. UNESCO Digital Library*. https://doi.org/.1037//0033-2909.I26.1.78

Hu, R. L., Granderson, J., Auslander, D. M., & Agogino, A. (2019). Design of machine learning models with domain experts for automated sensor selection for energy fault detection. *Applied Energy*, *235*(May 2018), 117–128. https://doi.org/10.1016/j.apenergy.2018.10.107

Hubert, M., Reynkens, T., Schmitt, E., & Verdonck, T. (2016). Sparse PCA for High-Dimensional Data With Outliers. *Technometrics*, *58*(4), 424–434. https://doi.org/10.1080/00401706.2015.1093962

Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, *47*(1), 64–79. https://doi.org/10.1198/004017004000000563

Irish Water. (2019). Strategic Funding Plan 2019-2024.

J., P. K., Singh, N., Dayama, P., & Pandit, V. (2019). Change Point Detection for Compositional Multivariate Data. Retrieved from http://arxiv.org/abs/1901.04935

Jackson, J. E., & Mudholkar, G. S. (1979). Control Procedures for Residuals Associated With Principal Component Analysis. *Technometrics*, *21*(3), 341–349. https://doi.org/10.1080/00401706.1979.10489779

Joe Qin, S. (2003). Statistical process monitoring: basics and beyond. *Journal of Chemometrics*, *17*(8–9), 480–502. https://doi.org/10.1002/cem.800

Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis.

Kouziokas, G. N. (2020). SVMkernelbasedonparticleswarmoptimizedvector.pdf.

Laory, I., Trinh, T. N., & Smith, I. F. C. (2011). Evaluating two model-free data interpretation methods for measurements that are influenced by temperature. *Advanced Engineering Informatics*, *25*(3), 495–506. https://doi.org/10.1016/j.aei.2011.01.001

Li, Wei, Peng, M., & Wang, Q. (2018). False alarm reducing in PCA method for sensor fault detection in a nuclear power plant. *Annals of Nuclear Energy*, *118*, 131–139. https://doi.org/10.1016/j.anucene.2018.04.012

Li, Wei, Peng, M., & Wang, Q. (2019). Improved PCA method for sensor fault detection and isolation in a nuclear power plant. *Nuclear Engineering and Technology*, *51*(1), 146–154. https://doi.org/10.1016/j.net.2018.08.020

Li, Weihua, Yue, H. H., Valle-Cervantes, S., & Qin, S. J. (2000). Recursive PCA for adaptive process monitoring. *Journal of Process Control*, *10*(5), 471–486. https://doi.org/10.1016/S0959-1524(00)00022-6

Lin, T.-K. (2018). PCA/SVM-Based Method for Pattern Detection in a Multisensor System. *Mathematical Problems in Engineering*, *2018*, 1–11. https://doi.org/10.1155/2018/6486345

Liu, X., Iftikhar, N., Nielsen, P. S., & Heller, A. (2016). Online Anomaly Energy Consumption Detection Using Lambda Architecture. In S. Madria & T. Hara (Eds.), *Big Data Analytics and Knowledge Discovery* (pp. 193–209). Cham: Springer International Publishing.

Liu, Y., Ma, X., Li, Y., Tie, Y., Zhang, Y., & Gao, J. (2019). Water pipeline leakage detection based on machine learning and wireless sensor networks. *Sensors (Switzerland)*, *19*(23), 1–21. https://doi.org/10.3390/s19235086

Maione, C., Barbosa, F., & Barbosa, R. M. (2019). Predicting the botanical and geographical origin of honey with multivariate data analysis and machine learning techniques: A review. *Computers and Electronics in Agriculture*, *157*(January), 436–446. https://doi.org/10.1016/j.compag.2019.01.020

Makaya, E., & Hensel, O. (2015). Water loss management strategies for developing countries: Understanding the dynamics of water leakages, 1–97. Retrieved from https://d-nb.info/1112580042/34

Mashford, J., De Silva, D., Burn, S., & Marney, D. (2012). Leak detection in simulated water pipe networks using SVM. *Applied Artificial Intelligence*, *26*(5), 429–444. https://doi.org/10.1080/08839514.2012.670974

Matos, C., Santos, C., Pereira, S., Bentes, I., Imteaz, M., Cook, S., … UNESCO. (2020). Environmental impact of water-use in buildings: Latest developments from a life-cycle assessment perspective. *Journal of Environmental Management*, *261*(2), 110198. https://doi.org/10.1016/j.jenvman.2020.110198

Moors, J., Scholten, L., van der Hoek, J. P., & den Besten, J. (2018). Automated leak localization performance without detailed demand distribution data. *Urban Water Journal*, *15*(2), 116–123. https://doi.org/10.1080/1573062X.2017.1414272

Moser, G., German Paal, S., & Smith, I. F. C. (2015). Performance comparison of reduced models for leak detection in water distribution networks. *Advanced Engineering Informatics*, *29*(3), 714–726. https://doi.org/10.1016/j.aei.2015.07.003

Mudry, A., & Tjellström, A. (2011). Historical background of bone conduction hearing devices and bone conduction hearing aids. *Advances in Oto-Rhino-Laryngology*, *71*, 1–9. https://doi.org/10.1159/000323569

Mujica, L. E., Rodellar, J., Fernández, A., & Güemes, A. (2011). Q-statistic and t2-statistic pca-based measures for damage assessment in structures. *Structural Health Monitoring*, *10*(5), 539–553. https://doi.org/10.1177/1475921710388972

Mutasa, S., Sun, S., & Ha, R. (2020). Understanding artificial intelligence based radiology studies: What is overfitting? *Clinical Imaging*, *65*(April), 96–99. https://doi.org/10.1016/j.clinimag.2020.04.025

Naderi, E., & Khorasani, K. (2016). A Data-driven Approach to Actuator and Sensor Fault Detection, Isolation and Estimation in Discrete-Time Linear Systems, *85*, 165–178. Retrieved from http://arxiv.org/abs/1606.06220

Nasir, M. T., Mysorewala, M., Cheded, L., Siddiqui, B., & Sabih, M. (2014). Measurement error sensitivity analysis for detecting and locating leak in pipeline using ANN and SVM. *2014 IEEE 11th International Multi-Conference on Systems, Signals and Devices, SSD 2014*, 7–10. https://doi.org/10.1109/SSD.2014.6808847

Nezhad, A. J., Wijaya, T. K., Vasirani, M., & Aberer, K. (2014). SmartD: Smart Meter Data Analytics Dashboard. *Proceedings of the 5th International Conference on Future Energy Systems*, 213–214. https://doi.org/10.1145/2602044.2602046

Nowicki, A., Grochowski, M., & Duzinkiewicz, K. (2012). Data-driven models for fault detection using kernel PCA: A water distribution system case study. *International Journal of Applied Mathematics and Computer Science*, *22*(4), 939–949. https://doi.org/10.2478/v10006-012-0070-1

Patabendige, S., Cardell-Oliver, R., Wang, R., & Liu, W. (2018). Detection and interpretation of anomalous water use for non-residential customers. *Environmental Modelling and Software*, *100*, 291–301. https://doi.org/10.1016/j.envsoft.2017.11.028

Pelz, G. (2003). *Mechatronic Systems*. *Library*. https://doi.org/10.1002/0470867906

Pérez, R., Sanz, G., Cugueró, M. À., Ramon, P., Sanz, G., & Angel, M.-. (2015). Parameter Uncertainty Modelling in Water Distribution Network Models Models, (September). https://doi.org/10.1016/j.proeng.2015.08.911

Perfido, D., Messervey, T., Zanotti, C., Raciti, M., & Costa, A. (2016). Automated Leak Detection System for the Improvement of Water Network Management. *Proceedings*, *1*(2), 28. https://doi.org/10.3390/ecsa-3-S5002

Posenato, D., Lanata, F., Inaudi, D., & Smith, I. F. C. (2006). Model free interpretation of monitoring data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4200 LNAI*, 529–533. https://doi.org/10.1007/11888598_47

Posenato, D., Lanata, F., Inaudi, D., & Smith, I. F. C. (2008). Model-free data interpretation for continuous monitoring of complex structures. *Advanced Engineering Informatics*, *22*(1), 135–144. https://doi.org/10.1016/j.aei.2007.02.002

Qin, S. J. (2012). Annual Reviews in Control Survey on data-driven industrial process monitoring and diagnosis, *36*, 220–234.

Qu, Z., Feng, H., Zeng, Z., Zhuge, J., & Jin, S. (2010). A SVM-based pipeline leakage detection and pre-warning system. *Measurement: Journal of the International Measurement Confederation*, *43*(4), 513–519. https://doi.org/10.1016/j.measurement.2009.12.022

Quevedo, J., Chen, H., Cugueró, M. À., Tino, P., Puig, V., García, D., … Yao, X. (2014). Engineering Applications of Arti fi cial Intelligence Combining learning in model space fault diagnosis with data validation / reconstruction : Application to the Barcelona water network. *Engineering Applications of Artificial Intelligence*, *30*, 18–29. https://doi.org/10.1016/j.engappai.2014.01.008

Quiñones-Grueiro, M., Bernal-de Lázaro, J. M., Verde, C., Prieto-Moreno, A., & Llanes-Santiago, O. (2018). Comparison of Classifiers for Leak Location in Water Distribution Networks ∗. *IFAC-PapersOnLine*, *51*(24), 407–413. https://doi.org/10.1016/j.ifacol.2018.09.609

Quinõnes-Grueiro, M., Verde, C., Prieto-Moreno, A., & Llanes-Santiago, O. (2018). An unsupervised approach to leak detection and location in water distribution networks. *International Journal of Applied Mathematics and Computer Science*, *28*(2), 283–295. https://doi.org/10.2478/amcs-2018-0020

Robles, D., Puig, V., Ocampo-Martinez, C., & Garza-Castañón, L. E. (2016). Reliable fault-tolerant model predictive control of drinking water transport networks. *Control Engineering Practice*, *55*, 197–211. https://doi.org/10.1016/j.conengprac.2016.06.014

Rosen, C. (2001). *A chemometric approch to process monitoring and control\n- with application to wastewater treatment operation-*. *Department of Industrial Electrical Engineering and Automation*.

Rousseeuw, P. J., & Hubert, M. (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(2), 1–14. https://doi.org/10.1002/widm.1236

Rousseeuw, P. J., Raymaekers, J., & Hubert, M. (2018a). A Measure of Directional Outlyingness With Applications to Image Data and Video. *Journal of Computational and Graphical Statistics*, *27*(2), 345–359. https://doi.org/10.1080/10618600.2017.1366912

Rousseeuw, P. J., Raymaekers, J., & Hubert, M. (2018b). A Measure of Directional Outlyingness With Applications to Image Data and Video. *Journal of Computational and Graphical Statistics*, *27*(2), 345–359. https://doi.org/10.1080/10618600.2017.1366912

Sabri, M., Sinal, B., & Kamioka, E. (2017). ADAPTIVE THRESHOLD BASED APPROACH TO PERFECTLY DETECT HEART CYCLE IN ECG DATA, (44), 492–498.

Saitta, S., Raphael, B., & Smith, I. F. C. (2005). Data mining techniques for improving the reliability of system identification. *Advanced Engineering Informatics*, *19*(4), 289–298. https://doi.org/10.1016/j.aei.2005.07.005

Salam, A. E. U., Tola, M., Selintung, M., & Maricar, F. (2015). Application of SVM and ELM Methods to Predict Location and Magnitude Leakage of Pipelines on Water Distribution Network, (19), 7970.

Samer El-Zahab. (2018). An accelerometer-based leak detection system _ Elsevier Enhanced Reader.pdf.

Sean Mulligan, Louise Hannon, Paraic Ryan, Sudeep Nair, E. C. (2020). Jo ur n re of. *Journal of Building Engineering*, 102248. https://doi.org/10.1016/j.janxdis.2020.102248

Sedki, A., & Ouazar, D. (2012). Hybrid particle swarm optimization and differential evolution for optimal

design of water distribution systems. *Advanced Engineering Informatics*, *26*(3), 582–591. https://doi.org/10.1016/j.aei.2012.03.007
Sengupta, R. N., & Kundu, D. (2016). Statistical Methods, 413–520.
Seyoum, S., Alfonso, L., Andel, S. J. Van, Koole, W., Groenewegen, A., & Van De Giesen, N. (2017). A Shazam-like Household Water Leakage Detection Method. *Procedia Engineering*, *186*, 452–459. https://doi.org/10.1016/j.proeng.2017.03.253
Sheriff, M. Z., Botre, C., Mansouri, M., Nounou, H., Nounou, M., & Karim, M. N. (2017). Process Monitoring Using Data-Based Fault Detection Techniques: Comparative Studies. *Fault Diagnosis and Detection*, (December). https://doi.org/10.5772/67347
Skworcow, P., Paluszczyszyn, D., Ulanicki, B., Jung, B. S., Boulos, P. F., Wood, D. J., … Giustolisi, O. (2013). Pressure , Leakage and Energy Management in Water Distribution Systems. *Water Resources Management*, *i*(1), 266. https://doi.org/10.1016/j.proeng.2015.08.855
Sliskovic, D., Grbic, R., & Hocenski, Z. (2012). Multivariate Statistical Process Monitoring. *Tehnicki Vjesnik-Technical Gazette*, *19*(1), 33–41. https://doi.org/10.1016/j.patcog.2009.11.008
Soldevila, A., Fernandez-Canti, R. M., Blesa, J., Tornil-Sin, S., & Puig, V. (2017). Leak localization in water distribution networks using Bayesian classifiers. *Journal of Process Control*, *55*, 1–9. https://doi.org/10.1016/j.jprocont.2017.03.015
Sousa, V., Silva, C. M., & Meireles, I. (2019). Performance of water efficiency measures in commercial buildings. *Resources, Conservation and Recycling*, *143*(October 2018), 251–259. https://doi.org/10.1016/j.resconrec.2019.01.013
Stavset, O., & Kauko, H. (2015). *Report -possibilities for smart energy solutions*.
Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., … Nenadic, G. (2019). Machine learning methods for wind turbine condition monitoring: A review. *Renewable Energy*, *133*, 620–635. https://doi.org/10.1016/j.renene.2018.10.047
Sydney Water. (2011). Best practice guidelines for water management in aquatic leisure centres.
Venkatasubramanian, V, Rengaswamy, R., Kavuri, S. N., & Yin, K. (2003). A review of process fault detection and diagnosis Part III: Process history based methods. *Computers & Chemical Engineering*, *27*(3), 293–311. https://doi.org/10.1016/S0098-1354(02)00160-6
Venkatasubramanian, Venkat, Rengaswamy, R., & Yin, K. (2003). A re v iew of process fault detection and diagnosis Part I : Quantitati v e model-based methods, *27*, 293–311.
Vieira, A. S., Beal, C. D., Ghisi, E., & Stewart, R. A. (2014). Energy intensity of rainwater harvesting systems: A review. *Renewable and Sustainable Energy Reviews*, *34*, 225–242. https://doi.org/10.1016/j.rser.2014.03.012
Villegas, T., Fuente, M. J., & Rodríguez, M. (2010). Principal component analysis for fault detection and diagnosis. experience with a pilot plant. *Proceedings of the 9th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics*, 147–152.
Vitter, J. S., & Webber, M. E. (2018). A non-intrusive approach for classifying residential water events using coincident electricity data. *Environmental Modelling and Software*, *100*, 302–313. https://doi.org/10.1016/j.envsoft.2017.11.029
Xiao, S., Lu, Z., & Xu, L. (2017). Multivariate sensitivity analysis based on the direction of eigen space through principal component analysis. *Reliability Engineering and System Safety*, *165*(March), 1–10. https://doi.org/10.1016/j.ress.2017.03.011
Xiao, W. (2016). A Probabilistic Machine Learning Approach to Detect Industrial Plant Faults. *International Journal of Prognostics and Health Management*, (c), 1–11.
Xu, L., Wang, X., Bai, L., Xiao, J., Liu, Q., Chen, E., … Luo, B. (2020). Probabilistic SVM classifier ensemble selection based on GMDH-type neural network. *Pattern Recognition*, *106*. https://doi.org/10.1016/j.patcog.2020.107373
Xu, X., Wang, H., Zhang, N., Liu, Z., & Wang, X. (2017). Review of the Fault Mechanism and Diagnostic Techniques for the Range Extender Hybrid Electric Vehicle. *IEEE Access*, *5*, 14234–14244. https://doi.org/10.1109/ACCESS.2017.2725298
Yu, Y., Woradechjumroen, D., & Yu, D. (2014). A review of fault detection and diagnosis methodologies on air-handling units. *Energy and Buildings*, *82*, 550–562. https://doi.org/10.1016/j.enbuild.2014.06.042
Zenobi, R., Knochenmuss, R., Yamashitat, M., Fenn, J. B., Wong, A., Jime, B., … Agency, E. (2011). Introduction to multivariate analysis. *J Am Chem Soc*. https://doi.org/10.1002/jms
Zhang, H., Qi, Y., Wang, L., Gao, X., & Wang, X. (2017). Fault detection and diagnosis of chemical process using enhanced KECA. *Chemometrics and Intelligent Laboratory Systems*, *161*(December 2016), 61–69. https://doi.org/10.1016/j.chemolab.2016.12.013
Zhao, Y., Li, T., Zhang, X., & Zhang, C. (2019). Artificial intelligence-based fault detection and diagnosis

methods for building energy systems: Advantages, challenges and the future. *Renewable and Sustainable Energy Reviews*, *109*(February), 85–101. https://doi.org/10.1016/j.rser.2019.04.021

Zhou, A., Yu, D., & Zhang, W. (2015). A research on intelligent fault diagnosis of wind turbines based on ontology and FMECA. *Advanced Engineering Informatics*, *29*(1), 115–125. https://doi.org/10.1016/j.aei.2014.10.001

Zumoffen, D., & Basualdo, M. (2008). From large chemical plant data to fault diagnosis integrated to decentralized fault-tolerant control: Pulp mill process application. *Industrial and Engineering Chemistry Research*, *47*(4), 1201–1220. https://doi.org/10.1021/ie071064m