Palau Estevan, CV.; Arregui De La Cruz, F.; Carlos Alberola, MDM. (2012). Burst detection in water networks using principal component anlysis. Journal of Water Resources Planning and Management. 138(1):47-54. doi:10.1061/(ASCE)WR.1943-5452.0000147.

BURST DETECTION IN WATER NETWORKS USING PRINCIPAL COMPONENT ANALYSIS

C.V. Palau [1], F.J. Arregui [2], M. Carlos [3]

[1] Associate Professor, Dept. of Rural Engineering. Hydraulic Div. Centro Valenciano de Estudios del Riego. Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain. E-mail: virpaes@agf.upv.es.

[2] Researcher, ITA. Universidad Politécnica de Valencia, Camino de Vera s/n. 46022 Valencia, Spain. Apart. 22012. E-mail: farrregui@ita.upv.es

[3] Associate Professor, Dept. Mechanical Engineering. Jaume I University, Av. de Vicent Sos Baynat s/n. 12071 Castelló de la Plana, Spain. E-mail: mcarlos@emc.uji.es

Abstract: The following work presents a multivariate statistical technique applied to the control of water inflows into district metering areas (DMA) of urban networks. This technique, called Principal Component Analysis (PCA), allows for a sensitive and quick analysis of the inflows into a DMA without hassling mathematical algorithms. PCA technique simplifies the original set of flow rate data recorded by the SCADA system, synthesizing the most significant information into a statistical model which is able to explain most of the behaviour of the water distribution network. PCA technique also allows for the establishment of control charts that help system operators in the identification of anomalous behaviours regarding water use, bursts or illegal connections. The described technique has been proved to offer high detection sensitivity to bursts or other unexpected consumptions.

Keywords: burst detection; principal component analysis; water distribution systems, real loss reduction.

INTRODUCTION

In the last decades water industry has invested a considerable amount of financial and human resources to reduce water losses in water networks. Numerous international initiatives like the Specialist Group on Water Loss of the International Water Association (IWA) are enriching industry knowledge on leak detection and location methodologies, contributing to the development of standardized procedures that allow water utilities to improve non revenue water (NRW) indicators. Despite these efforts, water losses in many countries around the world can be as high as 70% of the water input to the system. Under these circumstances it is clear that a considerable amount of work still needs to be conducted and simple methodologies, easy to implement in difficult environments like the ones found in underdeveloped or developing countries, need to be formulated.

Often, leakage problem comes as a result of many years of an inadequate investment in pipeline infrastructure, maintenance and replacement. As water distribution networks become older, pipes degrade and suffer frequent bursts and leaks which lead to unacceptable levels of water losses and, too frequently, to undesirable service disruptions. Other times, a deficient technical operation of the water network will be the reason for a poor infrastructure condition. This is the case when booster stations are not correctly protected against pressure surges, and the start-up and stop of the pumps cause significant pressure oscillations in the network.

High levels of water losses can be lowered if a larger number of occurring bursts and leaks are detected and the response time between pipe failure and repair is reduced. The total volume of water losses in bursts events depends on three factors: the magnitude of the leaks (flow rate), the burst occurrence rate and the time the water utility takes to identify and fix the problem.

Water network operators need to act upon all three factors to reduce water losses. Burst flow rate and incidence rate can be lowered by a proper pressure management of the network (UKWIR, 2003). The last one, the time a burst is active, depends on the effectiveness of water control practices. Burst duration, calculated from the time it first occurs until it is fixed, is the accumulated time of detecting, locating and repairing the leak (Water Research Centre, 1994). Location time

mainly depend on the availability, resources and qualification of the leak detection team. Quite often repair time has more to do with organizational issues and financial resources availability than with technical matters.

The method presented in this work aims at reducing the response time by helping network operators in the analysis of the data recorded by the telecontrol systems. The technique presented provides an excellent balance between its mathematical complexity, its easiness of use and the amount of information that can be extracted from the original data.

BACKGROUND

At present time water utilities have started the development of more or less elaborated strategies for the analysis of hydraulic variables within the network. The aim of these strategies, mostly based on statistical techniques, is to improve operator's response time and detection effectiveness against pipe failures. Such strategies have only been possible by means of real-time control and recording of the primary hydraulic variables using newly designed instrumentation with improved communication capabilities. Unfortunately, this enormous amount of recorded data has created a new problem. It needs to be processed, analysed and converted into useful information to allow for decision making activities.

Most of these methods of analysis allow researchers to define, on the basis of different factors, which should be the normal (expected) behaviour of the network. Then, when the reading of a sensor is too far from the expected value or it has an excessively large variation with respect the previous recorded value an alarm can be generated (Tsang, 2003). As a result of the research already conducted water operators have a number of these techniques available for leak detection in water distribution networks. ARIMA models, for example, exploit the property by which under normal conditions current inflows into the network are related to earlier flow values and variability. This technique can be either use to detect a fault or to reconstruct a flow pattern (Quevedo et al., 2010). Artificial neural networks (ANN) use continually updated historic data for creating a

probabilistic model of future flow profiles. The prediction of future flows allows for the identification of abnormal behaviours of water demand or burst events (Mounce and Machell, 2006; Mounce et al. 2010). Another approach in this field was proposed by Poulakis et al. (2003), a Bayesian probabilistic framework applied to flow data in water distribution systems for leak detection. Misiunas et al. (2006) presented a technique which allowed for the detection and location of bursts by a continuously monitoring of the inflows and pressure levels at different points of a DMA. The proposed method used the modified cumulative sum (CUSUM) test to locate the burst. Other methods of analysis found in technical literature that can be applied to water networks real time control and burst detection include: time-series analysis techniques (Prescott and Ulanicki, 2001), Kalman filters (Piatyszek et al. 2000), parity equations (Ragot and Maquin, 2006), pattern recognition methods (Valentin and Denoeux, 2001).

This paper describes a multivariate statistical technique, called Principal Component Analysis (PCA), for burst detection in urban water networks. It exploits the property by which the hourly flows can be condensed in a reduced number of variables which are calculated as a linear combination of those hourly flows. This method was previously applied by Palau C.V. (2005). The methodology creates different statistical models to simulate regular water demand establishing specific control charts that allow for the detection of abnormal operational conditions in the system.

PRINCIPAL-COMPONENT ANALYSIS METHODOLOGY

The PCA technique squeezes a high-dimensional data matrix (such as that conformed by the inflows to a DMA in a defined period of time) into a low-dimensional matrix in which the data variability is explained by a fewer number of variables called latent variables. Each observation (i.e. day of measurement) is stored in a different row of the matrix and is constituted of *K* measurements that are placed in the corresponding column of the row. An observation may be defined, for example, by the hourly (or any other time step) inflows into a DMA. The dimension of the original data matrix, denoted by $Z_{(N,K)}$, is defined by the number of days recorded, *N,* and the

number of measurements considered each day, *K.* Hence, element $x_{i,k}$ – i,k matrix position – correspond to i^{th} day and the k^{th} inflow measurement taken that day. This data matrix can be depictured as a data cluster of *N* points in a *K*-dimensional space.

$$Z_{(N,K)} = \begin{matrix} \text{Day } 1 \\ \vdots \\ \text{Day } N \end{matrix} \begin{bmatrix} Q_{1,1} & \cdots & Q_{1,K} \\ \vdots & \ddots & \vdots \\ Q_{N,1} & \cdots & Q_{N,K} \end{bmatrix} \quad (1)$$

Frequently, data compression techniques may lead to a significant loss of information. Thus it is crucial to find a compression technique which keeps as much information as possible from the original inflows. PCA reduces the original *K*-dimensional space of the inflows into a smaller *A*-dimensional subspace preserving the maximum amount of information. In other words, PCA transforms each measurement day, constituted by *K* inflow measurements, into a new simplified measurement day composed solely by *A* variables being each one a linear combination of the original *K* measurements. This new reduced space should still point up dominant patterns and major trends in the flow data.

Also, since several PCA models can be built for different time periods of the day, this technique allows for the establishment of various control charts in order to detect, in quasi real-time, pipe bursts or unusual water demands in the water network.

Before starting the transformation it should be noted that multivariate projection methods, like PCA, are sensitive to different numerical ranges of the variables (for example, hourly flow rates). An hour of measurement with a larger numerical range automatically gets more significance than an hour of measurement with a smaller numerical range. Therefore, the starting point when conducting a PCA, consists of a pre-processing step (Step 1 in Fig. 1) to transform the original inflow matrix $Z_{(N,K)}$, into the $X_{(N,K)}$ matrix by a transformation process called mean centering and scaling (Eriksson et al., 2001). For every flow rate, the average inflow at that time period is subtracted. Then flow rates are divided by the standard deviation of the flow rates at that particular time (Eq. 2).

$$X_{(N,K)} = \begin{bmatrix} \frac{(Q_{1,1} - \overline{Q_1})}{\sigma_1} & \cdots & \frac{(Q_{1,K} - \overline{Q_K})}{\sigma_K} \\ \vdots & \ddots & \vdots \\ \frac{(Q_{N,1} - \overline{Q_1})}{\sigma_1} & \cdots & \frac{(Q_{N,K} - \overline{Q_K})}{\sigma_K} \end{bmatrix} \quad (2)$$

Once this pre-processing step of the data is completed, PCA starts by calculating the eigenvector of the covariance matrix from the pre-processed $X_{(N,K)}$ matrix (Step 2 in Fig. 1). The computed vectors, called principal components (PC), define a reduced *A*-dimensional space which constitutes the PCA-model. The projections of the original measurements onto these PC generate new variables, called latent variables or scores, denoted by *t*. After the transformation, each measurement day can be defined by *A* latent variables (instead of the original *K* measurements) which are calculated as a linear combination of the original *K* inflows. The coefficients used to combine the *K* inflows to project them onto each PC are called loads. This linear transformation, to obtain the scores of every observation, $T_{(N,A)}$, can be easily done by multiplying the inflow matrix $X_{(N,K)}$ and the loading matrix $P_{(K,A)}$.

$$\mathbf{T_{(N,A)} = X_{(N,K)} \cdot P_{(K,A)}} \quad (3)$$

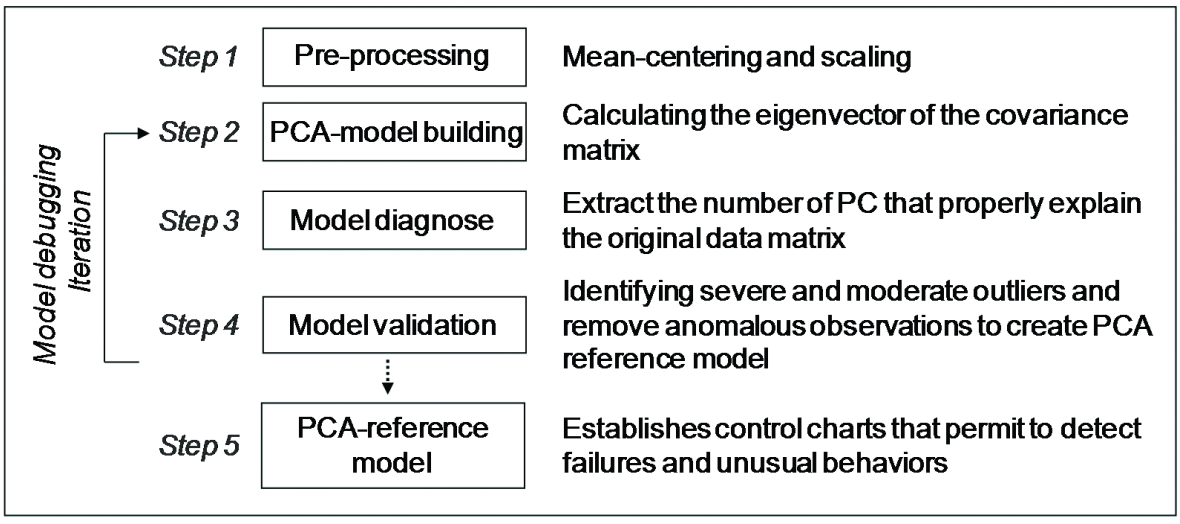


Fig 1. Methodology for PCA model building

Because of the reduction in space dimensionality from *K* to *A* variables, for every day and measured inflow there is a fraction of the inflow variability that the PCA model cannot explain. This

unexplained fraction of the data is represented by matrix, $E_{(N,K)}$. Thus, the original pre-processed data matrix $X_{N,K}$ is decomposed by PCA as:

$$\mathbf{X}_{(N,K)} = \mathbf{T}_{(N,A)} \cdot \mathbf{P}_{(K,A)}^{T} + \mathbf{E}_{(N,K)} \quad (4)$$

Where X` is the projected inflow matrix calculated by the PCA model, E the residual matrix, P the loading matrix and $T_{(N,A)}$ the score matrix. The score matrix, $T_{(N,A)}$, describes the orthogonal projections of the primitive data on the PC space. The P matrix or loading vectors define, in the original space, the directions that characterize the PC space. It is constituted by a number of vectors equal to the dimensionality of the new reduced space, *A*. The *A* loading vectors define the factors that multiply each measurement day to obtain the new *A* variables. Matrix $E_{(N,K)}$ contains the residuals or information that is not explained by the PCA-model (Wold et al., 1987). This matrix is a measure of how good the model fits the data and is also used to calculate acceptance/rejection statistics for future measurement days.

Another concern when building a PCA model is how to find the optimal number of principal components. On the one hand, reducing in excess the space dimensionality may cause a significant loss of information. On the other hand, extracting too many principal components leads to an over fitting of the model losing, its reliability and predictive capability. Therefore, it is essential to extract the correct number of principal components. Not too many, so data representation is not significantly simplified, and not too few, in which system behaviour is not satisfactorily explained by the PCA model.

Malinowski (1977, 1987) proposed different tools for determining the optimal PC space dimension and for diagnosing the model quality (Step 3 in Fig. 1). In general, it can be stated that the extraction of new principal components stops when adding a new variable does not significantly improve the explanation of the variables behaviour. Several parameters such as explained variation, R^2, which measures the “goodness of fit” or predicted variation, Q^2, that indicates the predictive capability of the model, are commonly used.

Frequently, not all the flow measurements extracted from the SCADA can be used for building the PCA model. For example, the days between the occurrence of a burst and its repair will distort the normal behaviour of the water network. The same happens when an abnormal consumption takes

place for a few days/hours. Obviously, these days should not be considered when building the reference model. Otherwise, these exceptional flow measurements will cause the model to accept a larger variability of the inflows within the normal behaviour of the DMA.

Therefore, building a model requires a formal procedure to detect and remove observations that introduce misleading flow variability. This process of identifying unwanted observations (outliers) that may distort the reference model is called model validation (Step 4 in Fig. 1). Outliers can be classified, into severe or moderate, depending on their effect on the PCA model (Jackson, 1991). Different statistical techniques can be used to detect each outlier type.

Severe outliers are those observations in the PC space whose distance to the centre of gravity of the data cluster is considered to be too high. Severe outliers can attract towards themselves principal directions, those of maximum data variability, creating a fictitious component and making real data variability misleading (Fig. 2a). In this case, the two encircled observations exhibit a large distance to the centre of gravity of the data cluster. These observations deviate the direction of the calculated PC, P*, due to the artificial variability that they generate in the data cluster. As seen in Fig. 2a, this PC space does not correctly reproduce the original data variability. In contrast, when the outliers are removed from the model the new calculated PC, P, reproduce, significantly better, the data variability.

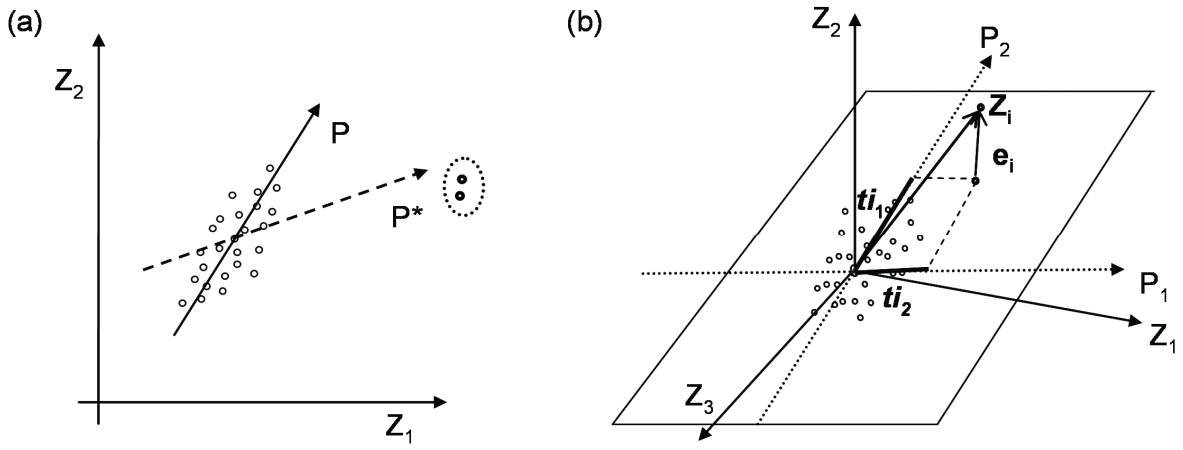


Fig 2. (a) Severe outlier; (b) moderate outlier

In the particular case of a PCA describing the measured inflows into a water network, a severe outlier appear if the flows measured are exceptionally high or low even in those cases in which the correlation structure of the flows is maintained. For example, a severe outlier can be caused by an observation in which the flow modulation curve moves in a vertical axis (all inflows increase or decrease by the same magnitude) and the shape of the curve is still similar to the average modulation curve. This will be the typical case in which a burst occurs before the analysed time period of the model starts.

Severe outliers are identified by the vectorial statistic T^2 Hotelling, following a F-Snedecor probability distribution with *A, N-A* degrees of freedom (Hotelling, 1947; Kourti and MacGregor, 1995). By means of this property it is possible to define control limits, with a specified confidence level, that will identify severe outliers. Typically this control limit, because of its vectorial nature, is presented in a plane in the form of an ellipse (Fig. 3).

Moderate outliers are those whose Euclidean distance to the model, or in other words the residual vector module $\|\mathbf{e}_i\|$, is exceptionally large, Fig. 2 (b). The statistical parameter used to identify this type of outliers is the Distance to Model (DMOD) defined by the ratio S_i / S_o, where S_i represents the absolute distance to the model, and S_o is the normalized distance to the model (Eriksson et al., 2001). The data needed to calculate S_o is contained in the residual matrix $E_{(N,K)}$.

The absolute distance of one observation divided by the normalized distance to the model squared, $(S_i/S_o)^2$, approximates a F-Snedecor probability distribution with *(K-A),(N-A-1)(K-A)* degrees of freedom. By means of this property it is possible to compute the membership probability of each observation.

A moderate outlier can be caused by an observation with flow values not necessarily too high or too low, in which the correlation structure between measured flows have been broken. Thus, DMOD statistic sensitively detects any change in the shape of the flow modulation curve. For example, an observation can be classified as a moderate outlier when a burst appears during the time period being analysed by the model. In such case the shape of the curve changes from the reference shape.

Model validation is an iterative process. Once outliers are detected and removed from the data matrix, the methodology returns again to the model building stage using the debugged data matrix. Obviously, additional outliers will always appear in the new model. These outliers have to be removed until they do not significantly distort the model. The iterative process finishes when the number of outliers approximately represent the percentage defined by the confidence value α. Then the reference PCA model is calculated and prepared to be put into operation in the SCADA system as a routine to detect failures and unusual water demands in the water network in real time (Fig. 1, step 5).

PCA leak detection principles

When analysing an observation in which a leak, having a flow rate Q_{leak}, is superimposed to the inflow, the projection of the total inflow, consumption and leak, T^{leak}, is displaced, in each PC direction, proportionally to the magnitude of the leak and the sum of the loads in that direction.

$$T^{leak} = \left[t_1^{leak} \quad \cdots \quad t_A^{leak} \right] = \left[\frac{((Q_{1,1} + Q_{leak}) - \overline{Q_1})}{\sigma_1} \quad \cdots \quad \frac{((Q_{1,K} + Q_{leak}) - \overline{Q_K})}{\sigma_K} \right] \begin{bmatrix} p_{1,1} & \cdots & p_{1,A} \\ \vdots & \ddots & \vdots \\ p_{K,1} & \cdots & p_{K,A} \end{bmatrix} \tag{5}$$

In particular, the score in a generic PC direction, i:

$$t_i^{leak} = \left[\frac{((Q_{i,1} + Q_{leak}) - \overline{Q_1})}{\sigma_1} \cdot p_{i,1} \quad \cdots \quad \frac{((Q_{i,K} + Q_{leak}) - \overline{Q_K})}{\sigma_K} \cdot p_{i,K} \right] = t_i + Q_{leak} \sum_{j=1}^{A} \frac{p_{i,j}}{\sigma_j} \tag{6}$$

Therefore, T^2 Hotelling statistic will detect leaks more efficiently in those directions with a greater sum of the loadings or when all loadings have the same sign. In case, the sum of the loadings is almost null, the T^2 Hotelling statistic in that direction will not be able to detect efficiently any leak that starts before the initial time analysed by the model. In case the leak starts after that time then the displacement will be proportional to the sum of the loadings from the leak occurrence time until the latest time considered by the model. In such case the effectiveness cannot be predicted since it will depend on the sign of loadings, their value and the variability of the observations used to build the model in that direction.

When analysing the effectiveness of the DMOD statistic things become slightly more complex. The residual generated by the leak will have to be added to the residual of the observation itself.

Therefore, the effect of the leak on the DMOD parameter will not necessarily be proportional to the flow rate. Nonetheless, in general terms, it can be stated that the residuals of the total inflow (leak plus consumption) will be greater if the leak starts after the initial time analysed by the model. In such case, the shape of the modulation curve changes and the model cannot properly fit the data and, consequently, DMOD parameter increases. Contrarily, when the leak starts before the time period analysed, the shape is not significantly altered and the DMOD parameters suffers a negligible change.

CASE STUDY

District Metering Area Description

In order to evaluate the effectiveness of the PCA analysis technique, a PCA model has been built of the inflows into a DMA belonging to a municipal water utility in the eastern coast of Spain. The water is supplied by gravity into the distribution network from an elevated reservoir at the top of a hill. The water network is divided in six independent DMAs which are monitored by a SCADA system. In all DMAs flows were measured by mechanical Woltmann meters, of different sizes, equipped all of them with pulse emitters having a resolution of 100 l. Flow data into the different DMAs are recorded in local controllers in five minutes intervals. The controllers send, by means of radio transceivers, the average flow every hour to the central server. This is the value finally recorded in the SCADA database.

The city has been partly built on a hill creating large pressure differences within the water network. This problem is also observed in the DMA under study where large pressure variations are found. Unsurprisingly, pressure in areas close to the reservoir at the top of the hill is relatively low, close to 2 bar, while in the lower areas can become as high as 6 bar. The DMA studied in the present work consists of approximately 2900 residential properties with some commercial and industrial water users which approximately represent 15% of the water use.

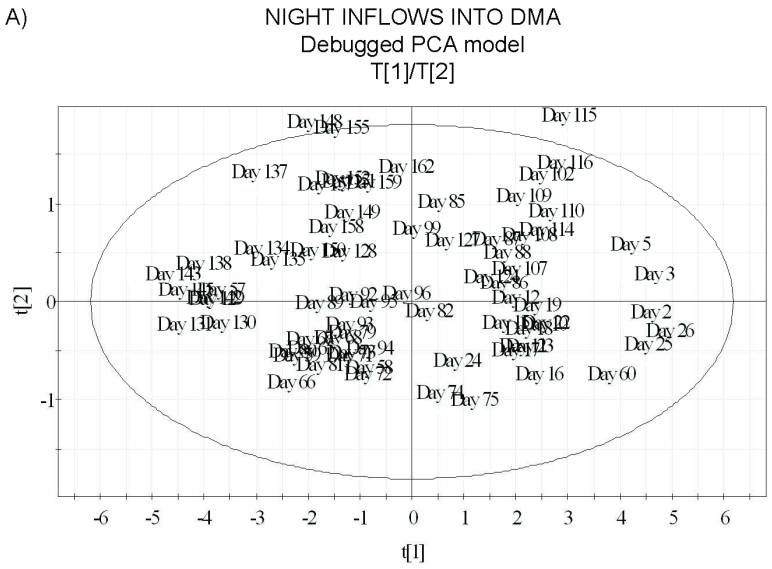
In general terms, pipe infrastructure was in poor maintenance conditions and there was frequent burst occurrence and a high natural rate of rise of leakage (Lambert, 2001), as shown in Fig. 4, which led to numerous repair interventions in the network. This characteristic of the DMA, without

any doubt, made more challenging the establishment of valid references for the inflows. Obviously, a more stable DMA will allow setting stricter control limits and a better detection effectiveness of the method.

Building Principal-Component Analysis Models

For the particular case study presented a six month period of flow data, from January until June 2007, was extracted from the SCADA system. Flows were recorded in hourly terms and expressed in litres per second. According to the water company the operational conditions of the DMA were stable during the period considered. No large pressure or water demand variations due to seasonal effects or population increases were recorded. Taking into account the whole period, the average daily inflow did not show a dependence on time, i.e. total daily inflows can be considered a stationary time series.

A PCA model was built for different time periods in order to achieve a better reaction time for burst detection. Building different models throughout the day increases data homogeneity and reduces the number of principal components needed in each model, facilitating the analysis procedures and the use of control charts (Fig. 3). In fact, one model which considered the complete day (each observation was constituted by 24 hourly flows) was also built. However, the model required five PC components to properly describe data variability (Table 1).



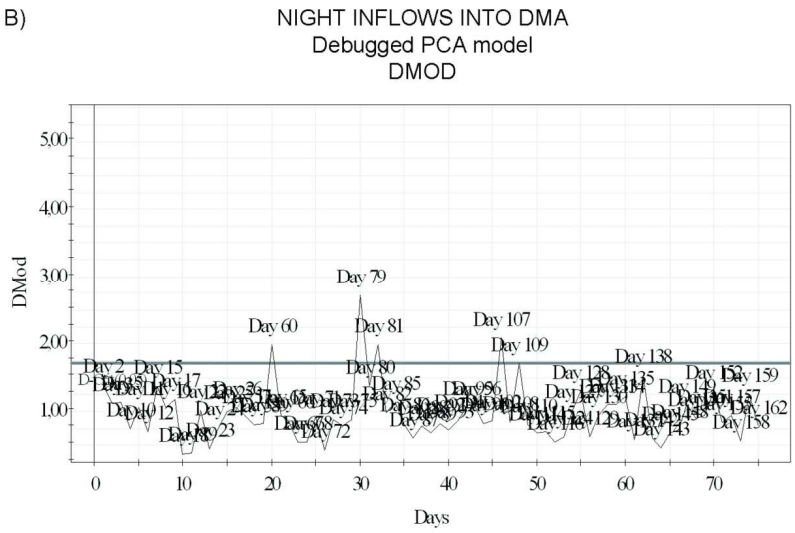


Fig 3. Debugged PCA model control charts: (A) T^2 Hotelling (B) distance to model; confidence level 95%; R^2 (cumulative)= 95.2%; Q^2 (cumulative) = 89.5%.

Table 1. PCA models description

Model	*Number of PC*	*Eigen value*	*R2 (% cum.)*	*Q2 (% cum)*	*Nº of observ.*
Night	1	6.14	87.7	83.4	75
	2	0.52	95.2	89.5	
Morning	1	4.31	47.9	36.4	68
	2	2.15	71.8	51.4	
	3	1.12	84.3	58.5	
Afternoon	1	6.22	77.7	70.2	89
	2	1.09	91.4	84.7	
24 h	1	17.4	72.5	71.0	69
	2	2.07	81.2	72.3	
	3	1.81	88.7	80.8	
	4	0.93	92.6	83.3	
	5	0.66	95.3	87.8	

A total of 180 observations were available for building three PCA models covering the following time periods: from 12:00 a.m until 6:59 a.m. (night), from 7:00 a.m. until 15:59 p.m (morning) and from 16:00 p.m until 23:59 p.m (afternoon). The behaviour of the network was also found to be different between working days and weekends (Fig. 5). Again, in order to reduce unnecessary variability of the data which will decrease the detection effectiveness of the methodology different models were prepared for working days (130 days) and weekends (50 days). As seen in Fig. 5, this distinction is particularly important for morning and afternoon models.

For building the models, all steps described in Fig. 1 were followed until the optimum number of PC was reached and all significant outliers removed. Night flow data matrix was reduced from 7 to 2 variables, morning flow matrix from 9 to 3 and afternoon flow matrix from 8 to 2. Also, after outlier detection, a significant number of observations from the initial 130 working days had to be discarded to build the models. The explanation for this high percentage of rejected observations can be found in the high instability of the system, as shown in Fig. 4.

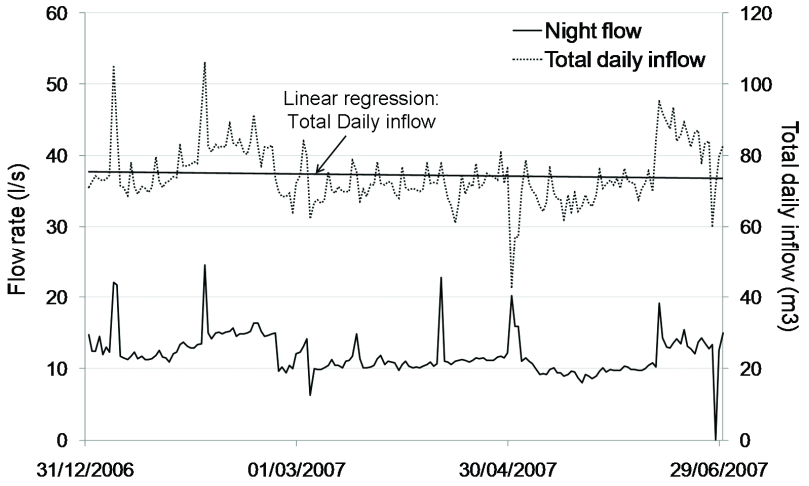


Fig 4. Night flow and total inflows into the studied DMA.

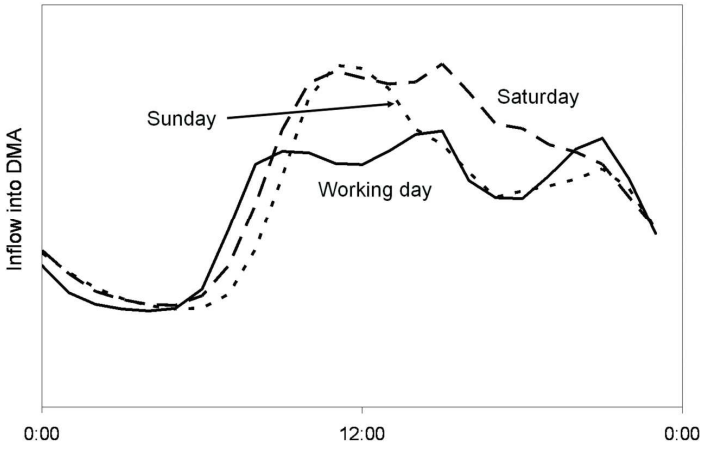


Fig 5. Average water flow pattern during working days and holidays.

For each model, a T^2 Hotelling and DMOD control charts were prepared in order to have a visual identification of future outliers as a consequence of bursts, exceptional water uses or any other anomalous behaviour in the water network. As an example, Fig. 3 shows T^2 Hotelling and DMOD control charts resulting from the PCA model developed for night flows measured during the

considered working days. This figure also shows the location in the chart of the observations finally used to build the model.

Once the models have been built every new day of measurement corresponds to a new observation. For every new observation, outlier detection is mathematically carried out by calculating the two statistics previously presented: DMOD and T^2 Hotelling. Observations which projections lay out of the T^2 Hotelling control limits (defined by an ellipse) are considered severe outliers, while points out of DMOD control limit are defined as moderate outliers. Severe outliers typically show inflows which are either too high or too low. Moderate outliers are days in which a burst or an abnormal consumption has occurred during the time interval being analysed.

Finally, it should be mentioned that PCA technique, if properly used, has the ability of classifying faulty observations into different groups: burst, service interruptions and abnormal consumptions. Each one of these abnormalities will produce a different deviation or effect on the projections onto the PC space. For example, as it has been seen in Eq. 6, a burst will displace the projections proportionally to the flow rate of the leak and the sum of the loads in each PC direction. In this case, the DMOD parameter will not increase significantly (depending on the occurrence time of the burst). A service disruption will alter the projection onto the PC space and increase the DMOD parameter of the observation.

The time of occurrence of a burst or any other abnormal event can only be estimated if several successive models are built, and T^2Hotelling and DMOD statistics are properly used.

Principal-Component Analysis model evaluation by means of simulated consumption profiles

When analysing the performance of statistical detection methods it is important to bear in mind that faults detection effectiveness cannot be generalized in a single figure. Their effectiveness will always depend on the data variability used to build the models. In this particular case study, due to the high instability of the network, data variability of the flows used to build the models is extremely high. As a result, defining a reference behaviour of the system has been a difficult task and many observations, that clearly distorted the model, had to be rejected. Under these circumstances,

control limits are much less strict than in a DMA behaving in a very stable manner. A more settled DMA could have been chosen for the analysis, obtaining much more attractive results. However, it is important to show that this methodology do not only work under ideal conditions but also under not such favourable circumstances. In any case, in a situation like the one presented, PCA models (as any other statistical method) would have to be updated every now and then to adapt the models to the new behaviour of the network in order to achieve better detection effectiveness and reduce false alarms. This is where a relatively simple methodology, like the PCA, takes advantage over more complex techniques that will also have to be updated.

Additionally, all statistical detection methods are subject to two different types of error: a) False positives, also called α or Type I errors, which produce false alarms; b) False negatives, also called β or Type II errors, which will lead to the acceptance of a faulty observations, for example a flow modulation curve with a significant leak. If control limits are set too strict, using lower confidence levels (for example 75%), Type II errors will be avoided at the expense of producing frequent Type I errors, or false alarms. On the contrary, if control limits are set too loose selecting a high confidence level, Type I errors rarely occur but too many observations can be accepted having a significant leak. Technical managers have to find equilibrium between both types of errors to attain the maximum benefit from the statistical methodology.

For the analysis of the PCA model effectiveness thirty thousand random flow modulation curves have been generated. This approach, although constrained to the particular case study presented, allows for a more formal and controlled testing of the methodology than limited field surveys. Simulated flow curves have been generated by means of the property by which the residuals of the PCA model can be considered a Gaussian noise, i.e. residuals are distributed in each PC direction according to a Normal probability distribution having a zero mean and a standard deviation equal to the square root of the Eigen value in that direction (Table 1).

Afterwards, leaks of increasing magnitudes occurring at different times are added to the randomly generated inflows (Palau et al., 2004). T^2 Hotelling and DMOD control charts are then used to decide whether the resulting flow modulation curves are classified as outliers or not. The results of this performance analysis can be summarised in charts (Fig. 6) in which the detection

effectiveness is plotted. These graphs represent the effectiveness of T^2 Hotelling and DMOD control charts for the night model depending on the confidence levels selected. Two confidence level values, 75% and 95%, have been used for the calculation. Each line show how the occurrence of a leak at a specific time is detected with higher probability as it increases in flow (expressed as a percentage of the average flow during the time interval considered by the model).

Several conclusions are drawn from the analysis of the DMOD parameter effectiveness charts in Fig. 6. In first place, as expected, this statistic is significantly less effective if the leak occurs before the analysis period, i.e. before or at 12:00 a.m. in the night model, before or at 7:00 a.m. in the morning model or 4:00 p.m. in the afternoon model. The higher effectiveness is reached when the burst occurs in the middle of the time interval being analysed (for example 4:00 a.m. in the night model). At that time the change in shape of the curve is more evident and the DMOD parameter easily detects the burst. Fig. 6 also shows how increasing the confidence level decreases the effectiveness. In other words, Type I errors, or false positives are reduced at the expense of leaving some burst undetected, i.e. accepting faulty observations (Type II errors).

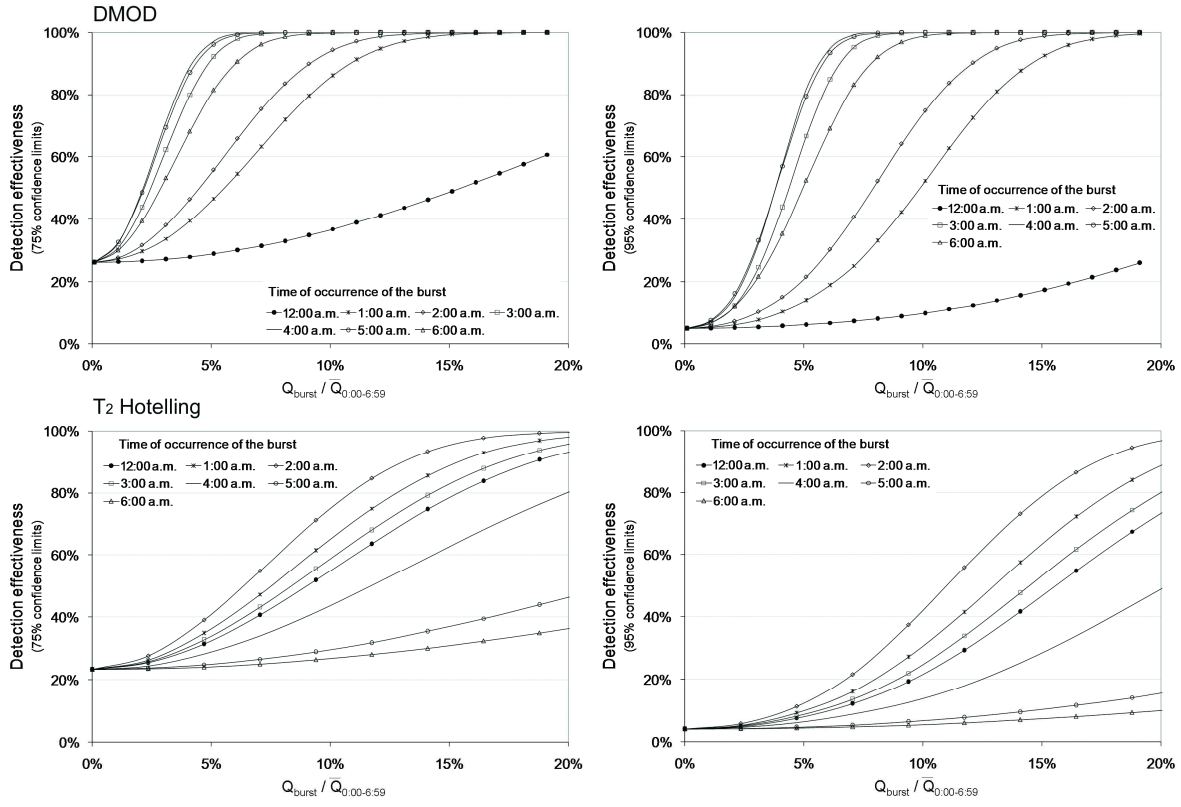


Fig 6. T^2 Hotelling and DMOD effectiveness for night flow model.

Similarly to the DMOD effectiveness chart the T^2 Hotelling statistics exhibits a different performance depending on the time of occurrence and the magnitude of the burst. However, this parameter mainly detects changes in the total consumption registered during the time interval being analysed. As a result, the detection effectiveness decreases as the burst occurs later in the time interval considered. In other words, T^2 Hotelling is more effective when the leak occurs before the flows are analysed or at the beginning of the time interval.

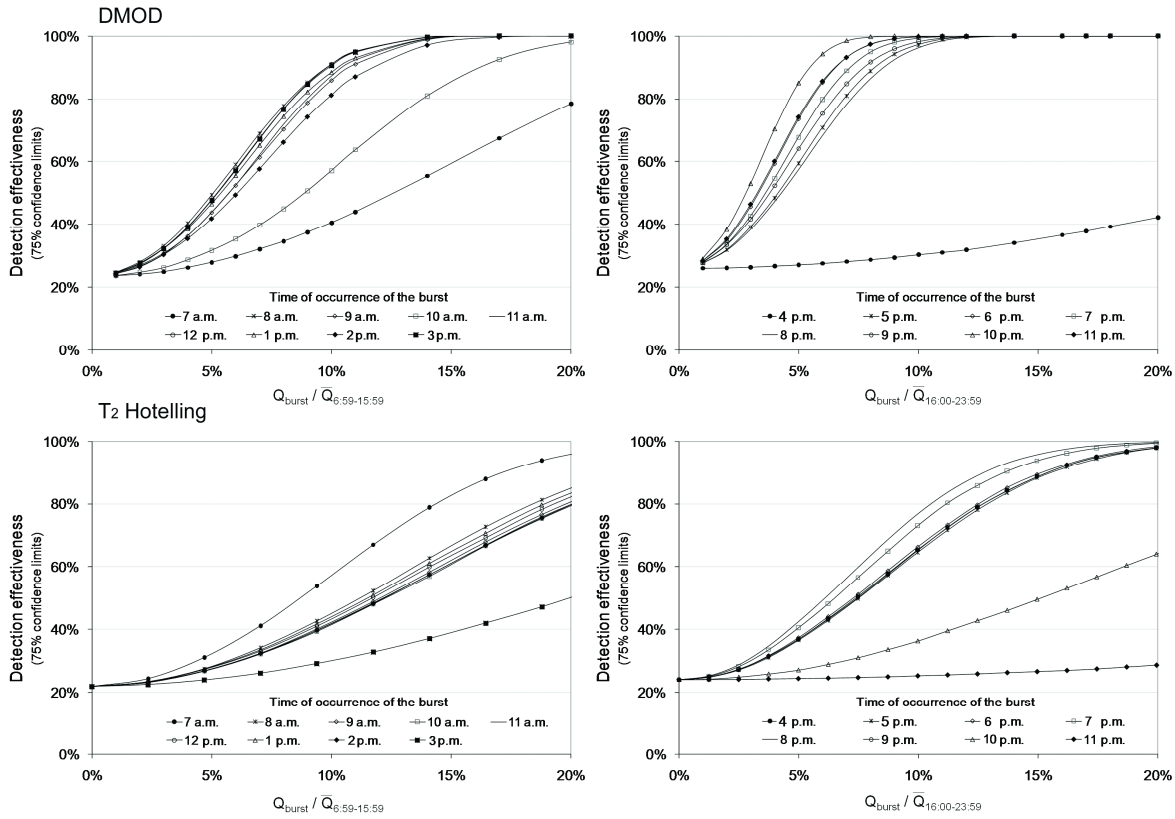


Fig 7. T^2 Hotelling and DMOD effectiveness for morning and afternoon flow models.

Fig. 7 shows the effectiveness charts of DMOD and T^2 Hotelling statistics calculated with a 75% confidence level for the morning and afternoon models. While the effectiveness of T^2 Hotelling is more or less the same for the three models, the DMOD statistic does not reach the high levels of effectiveness achieved in the night model.

When comparing both detection methods, is clear that bursts are better detected by studying the shape of the flow curves (DMOD) rather than their variability (T^2 Hotelling). This statement can also be checked by considering the effectiveness of the three models built: night, morning and afternoon (Fig. 8). The detection effectiveness of a leak representing 5% of the average flow is

always higher when using DMOD statistic. Also, the effectiveness of the night PCA model is significantly higher than the morning and afternoon PCA models, for which the variability of flows expand the control limits.

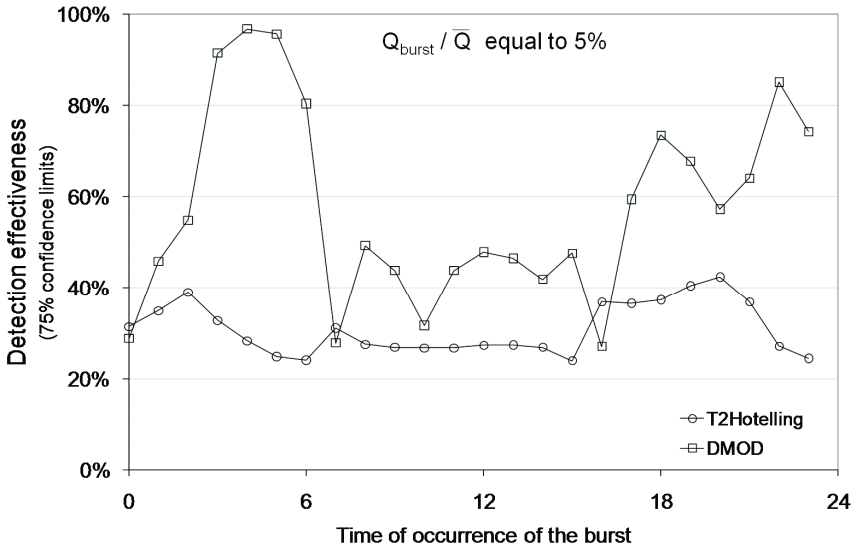


Fig. 8. Detection effectiveness as a function of its time of occurrence.

CONCLUSIONS

The basis of the methodology presented in this paper is founded in the geometry of the original data cluster that in this case corresponds to the water inflows into a DMA. The PCA model defines statistical control limits for T^2 Hotelling and DMOD, which allow for the detection of burst or abnormal behaviours of the system.

The final sensitivity of the presented technique strongly depends on the quality and variability of the data used to build the model. For that reason, is highly advisable to stratify daily inflows in a) working/weekend days; b) different hourly intervals. By doing so, data variability is considerably reduced and a better burst detection effectiveness is achieved.

However, from the two statistics used to detect outliers DMOD has shown to have the best sensitivity against bursts or other incidences in the DMA. This parameter detects when the correlation structure between variables have been broken, in other words when the flow modulation curve changes in shape. In the tested DMA, regardless of its high instability, a burst of

approximately 5% of the average flow could be detected with a probability between 30% and 95% depending on the hour of occurrence.

REFERENCES

Eriksson, L., Johansson, E., Kettaneh-Wold, N, and Wold, S. (2001). "*Multi- and Megavariate Data Analysis Part I: Basic Principles and Applications*". Umetrics Academy Book. Umea, Sweden.

Hotelling, H. (1947). "*Multivariate quality control, illustrated by the air testing of sample bombsights*". Techniques of Statistical Analysis, McGraw-Hill. New York., 113-184.

Jackson, J.E. (1991). "*A user's guide to Principal Components"*. Wiley. New York.

Kourti, T and MacGregor J.F. (1995). *"Process analysis, monitoring and diagnosis using multivariate projection methods"*. Chemometrics Intell. Lab. Syst.. 28(1), 3 - 21.

Lambert, A. (2001). *What do we know about pressure: leakage relationships?.* Proc. IWA Conf. "System Approach to Leakage Control and Water Distribution Systems Management", Brno.

Malinowski, E.R. (1977). "*Determination of the number of factors and the experimental error in the data matrix*". Anal. Chem. 49 (4), 612-617.

Malinowski, E.R. (1987). "*Theory of the distribution of error eigenvalues resulting from PCA with applications to spectroscopic data*". J. Chemometr. 1 (1), 33-40.

Misiunas, D., Vítkovský, J., Olsson, G., Lambert, M., and Simpson, A. (2006). "*Failure monitoring in water distribution networks*". Water Sci. and Technol., 53 (4-5), 503–511.

Mounce, S.R., and Machell, J. (2006). "*Burst detection using hydraulic data from water distribution systems with artificial neural networks*". Urban Water J., 3 (3), 21-31.

Mounce S.R., Boxall, J.B., and Machell, J. (2010). "*Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows"*. J. Water Resour. Plann. Manage., 136(3), 309-318.

Quevedo, J., Puig, V., Cembrano, G., Blanch, J., Aguilar, D., Benito, G., Hedo, M., and Molina, A. (2010). "*Validation and reconstruction of flow meter data in the Barcelona water distribution network".* Control Eng. Practice. 18, 640-651.

Palau C.V., Arregui,F.J., and Ferrer, A. (2004). *"Using multivariate Principal Component Analysis of injected water flows to detect anomalous behaviors in a water supply system - a case study".* Water Sci. and Technol. Water Supply. 4 (3), 169-181.

Palau C.V. (2005). PhD Thesis. "*Aportaciones a los sistemas de medición de caudal en redes de distribución de agua a presión*". Universidad Politécnica de Valencia. Spain.

Piatyszek, E., Voignier, P., and Graillot, D. (2000). *"Fault detection on a sewer network by a combination of a Kalman filter and a binary sequential probability ratio test".* J. Hydrol., 230 (3–4), 258–268.

Poulakis, Z., Valougeorgis, D., and Papadimitriou, C. (2003). "*Leakage detection in water pipe networks using a Bayesian probabilistic framewor*k". *Probab. Eng. Eng. Mech.*. 18, 315-327.

Prescott, S.L., and Ulanicki, B. (2001). "*Time series analysis of leakage in water distribution networks*" In Water Software Systems-Theory and applications, vol.2. Research Studies Press Ltd. England.

Ragot, J., and Maquin, D. (2006).*"Fault measurement detection in an urban water supply network"*. J. Process Control, 16(9), 887–902.

Tsang, K.M. (2003).*"Sensor data validation using gray models".* ISA Transactions, 42, 9–17.

UK Water Industry Report, UKWIR (2003). "*Leakage index curve and the longer term effects of pressure management"*. UKWIR Report 03/WM/08/29. London. United Kingdom.

Valentin, N.,and Denoeux,T.(2001)."*A neural network-based software sensor for coagulation control in a water treatment plant"*. Intell. Data Anal., 5, 23–39.

Water Research Centre (WRc). (1994). "*Managing leakage*" Rep. A. U.K. Water Ind. Research Ltd.

Wold, S., Esbensen, K., and Geladi, P. (1987). "*Principal Component Analysis*". Chemometrics Intell. Lab. Syst., 2 (1-3), 37-52.