USING MULTIVARIATE PRINCIPAL COMPONENT ANALYSIS OF INJECTED WATER FLOWS TO DETECT ANOMALOUS BEHAVIORS IN A WATER SUPPLY SYSTEM. A CASE STUDY.

PALAU, C.V.[1] , ARREGUI, F.[1], FERRER, A.[2]

[1] Instituto Tecnológico del Agua. Organismo Público Valenciano de Investigación- Universidad Politécnica de Valencia. Camino de Vera s/n. Apartado 22012. 46071 Valencia. Spain

[2] Departamento de Estadística e Investigación Operativa Aplicadas y Calidad. Universidad Politécnica de Valencia. Camino de Vera s/n. Apartado 22012. 46071 Valencia. Spain

E-mail: virpaes@gmf.upv.es, farrregui@gmf.upv.es, aferrer@eio.upv.es

Abstract -. The amount of data collected by the SCADA of an urban water supply system is sometimes difficult to process. A multivariate statistical technique, Principal Component Analysis (PCA) is presented in this paper, which processes this data, simplifying and synthesizing the most significant information. This technique extracts new variables, principal components (PC), that explain the behaviour of injected flow. Multivariate control charts to detect outliers show higher sensitivity than those generated with traditional univariate statistical methods.

Keywords: PCA, control charts, injected flow, MSPC, leakage detection.

INTRODUCTION TO PRINCIPAL COMPONENT ANALYSIS (PCA)

Frequently, the amount of data collected by the SCADA system (Supervisory Control And Data Acquisition) in a water utility exceeds the human analysis capacity, since it continuously records great quantity of data of different variables. In most cases, these variables are related to each other, i.e., they are not independent. Since classical univariate statistical methods only apply when variables are uncorrelated and, in the studied case the water consumption measured during one hour is related to previous water demand, a multivariate technique should be used.

From a statistical point of view and in a restrictive way, multivariate analysis can be defined as “a group of techniques whose objective is the descriptive analysis and/or to build inferences from multivariate data, that is to say, in which each observation is constituted by the values of several interrelated variables” (Romero,1997). For this reason, multivariate statistical techniques are very useful to analyse great quantities of data and convert it into usable information. Multivariate theoretical principles are known for long time and have been developed and used in large number of areas such as Sociology, Medicine, Biology or Hydraulic, as the case study. There is an extensive classification of the different multivariate analysis techniques depending on the data matrix studied and the aim of the analysis.

In this paper, one of these multivariate methods, Principal Component Analysis (PCA), is used to study the injected night flows to a water distribution system. In this technique, variables constitute an homogeneous group and information is simplified into a new reduced number of variables, called latent variables or principal components (PC). These new variables, those which better explain the original data, are built as a linear combination of the original ones (Jackson, 1991). The correlation structure of the original variables is an essential aspect of this analysis, that uses the relationship between variables to achieve a deep, simple, and complete study of the observed data.

AIM OF PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal Component Analysis (PCA) is a statistical technique that compresses a high-dimensional data matrix into a low-dimensional subspace, in which most of data variability is explained by a fewer number of latent variables.

Commonly, original data are organized in a data matrix denoted by **Z**. Each object or observation,- is stored in a different row and is composed of *K* variables-, that are placed in columns. Matrix dimension is defined by the number of observations, *N*, and the number of variables *K*. Hence, element $x_{i,k}$ –i,k matrix position- is i^{th} measurement of the k^{th} variable. Data matrix can be depictured as a data cloud of *N* points in a *K*-dimensional space (Figure 1).

When the original data matrix is simplified some information is lost. Consequently it is important to find a new *A*-dimensional space, of principal components (PC), which is able to explain in the best possible manner the cloud of points of the original data. In this sense, PCA is a technique that allows to reduce the original *K*-dimensional space into a new *A*-dimensional subspace preserving the maximum information from the initial data. This new PC-space should point up dominant patterns and major trends in the data. It will also help to detect outliers, i.e. points with unusual behaviours. At the same time, PCA can be used to build a model describing how the system behaves and indicating most remarkable process variables.

For illustration, a simple example is shown in Figure 1. The original data cloud is plotted in a two-dimensional space. By means of the PCA technique, a one-dimensional space, which best describes the original data, is obtained. This direction is calculated so that the distance between the data points and their projections onto the one-dimensional space is minimised.

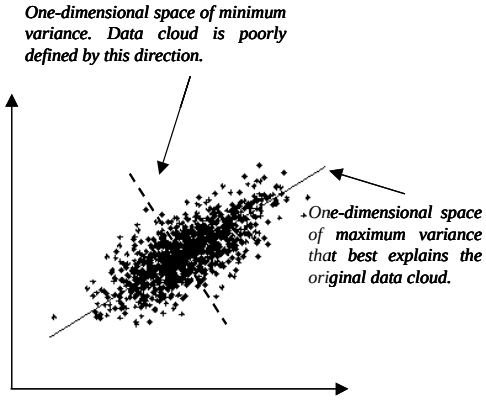

Figure 1.Data cloud projected into a line by PCA.

METHODOLOGY

The starting point in PCA, and in general in every multivariate analysis procedure, consists of finding an adequate data matrix and transform it into the best suitable form. This pre-processing is essential when building a mathematical model and can make the difference between a successful and an unsuccessful analysis.

Firstly, it should be noted that multivariate projection methods, like PCA, are sensitive to different numerical ranges in the variables. A variable with a large numerical range automatically gets more significance than a variable with a small numerical range. For this reason, the most common pre-processing tool involve a two step procedure. In the first step, for every observation, the average value of each variable is subtracted. In the second step, all the data is divided by the standard deviation of the corresponding variable. In this way, **Z** matrix is transformed into **X** matrix.

PCA analysis starts after data pre-processing (Figure 2), calculating the eigenvector of the covariance matrix of pre-treated **X**, during *model* building. The computed vectors, called principal components, define a

new *A*-dimensional space. The projection of the original objects onto these new principal directions generates new variables (latent variables or scores), linear combination of the original ones, that are uncorrelated and contain the most important information of the primitive *K*-variables. The new *A*-dimensional space constitute the PCA-model. As said before, and because of the reduction in space dimensionality from *K* to *A* variables, some variability of the data is not "explained" by the model. The fraction of the data that the PCA-model is no able to "explain" is considered to be statistical noise.

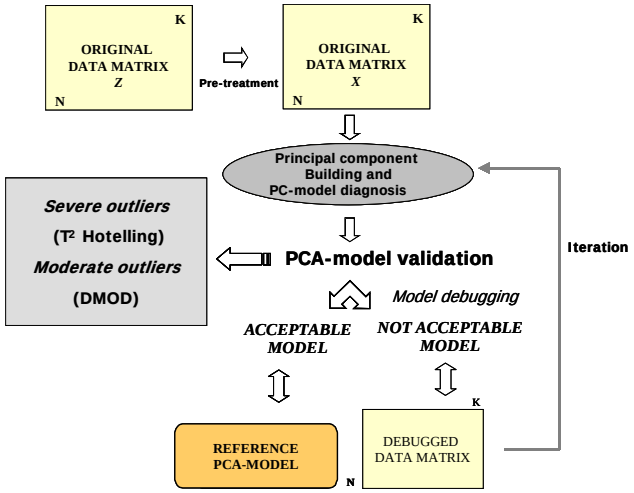

Figure 2. PCA-model building methodology.

Thus, the original pretreated data matrix $\mathbf{X}_{N,K}$ is decomposed by PCA as

$$\mathbf{X}_{(N,K)} = \mathbf{X}`_{(N,K)} - \mathbf{E}_{(N,K)} = \mathbf{T}\cdot\mathbf{P}^{T} + \mathbf{E}_{(N,K)} \quad (1)$$

Where

X: Original pretreated data matrix.
X`: Estimated data matrix calculated by PCA.
E: Residual matrix.
P: Loading matrix
T: Score matrix.

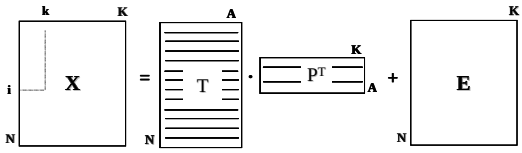

Figure 3. Decomposition of original data matrix X.

As seen in equation 1, **P** and **T**, are matrices of lower dimension which capture the essential patterns and trends of the observations. The **P** matrix or *loading* vectors defines the directions of principal components which constitute the new PC-subspace. Matrix **T** or *score* matrix describes the orthogonal projection of the primitive data on the PC-space (Figure 3). Matrix **E** contains the residuals, i.e. the information that is not "explained" by the PCA-model (Wold et al., 1987).

For finding those "lines and planes of closest fit to systems of points in space", that is the principal components or directions, a mathematical algorithm called NIPALS is used (Wold,1966).

Once the PC-space is defined, each observation can be decomposed into two other vectors (Figure 4). The first one is the orthogonal projection of the data point on the PC-space, **t**. The second one is the residual, i.e., the distance between the data point and its projection onto PC-space, **e**. As previously said, the later vector represents the amount of information that PCA-model is not able to reproduce.

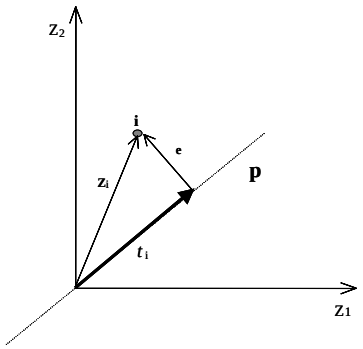

Figure 4. Vector decomposition on the new PC subspace.

A common concern in PCA is how to find the optimal number of principal components. On the one hand, reducing in excess the space dimensionality may cause significant information loses. 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 analysis 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 PC-space dimension and *diagnose* the model quality. In general, it can be said that the extraction of new principal components stops when adding a new variable is not significant and the model does not explain in a better way the behaviour of the variables. In technical literature several parameters, such as R^2, that measures the "*goodness of fit*" or Q^2 that indicates the predictive capability of the model (Eriksson et al., 1999) are already presented. For example, *Cross-validation* (CV) procedure was first developed by Wold (1978), and is a practical and reliable method to test the significance of a PCA model.

Up to this point, every observation stored in the data matrix has been used to build up the model. However, some of them, because of their different behaviour, may distort the directions of the principal components. For this reason, it is needed a *model validation* (Figure 5), that detects observations, called *outliers*, which may falsify the principal components space.

Outliers can be classified, into severe or moderate, depending on their effect on the PC model. Each type is detected using different statistical parameters.

- *Severe outliers* are those observations in the PC-space whose distance to the centre of gravity of the data cloud is considered to be too high. Severe outliers can orient towards themselves principal directions, those of maximum data variability, creating a fictitious component and misleading the real data (Figure 5a).

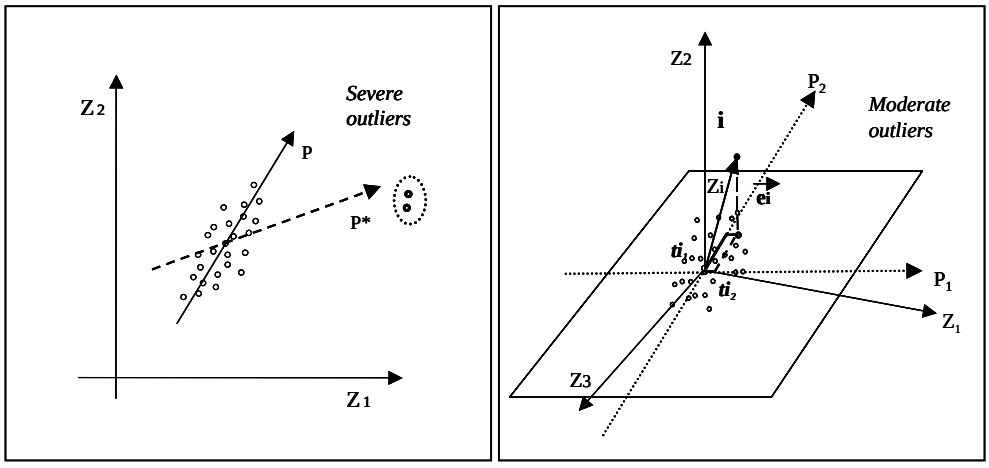

Figure 5. (a) Severe outlier, (b) Moderate outlier

For example, as depicture in Figure 5a, the distance of encircled observations to the centre of gravity of the data cluster is high. These observations may deviate the direction of the principal components. In this case, due to the fictitious variability that they cause, a direction p*, which does not correctly reproduce data variability, is extracted. Severe outliers mislead the PCA model due to the great effect that they exert during model building. *Model validation* (Figure 2) attempts to remove from the data matrix these dangerous outliers and then, principal directions are recalculated.

In practice, severe outliers are detected by the statistic T^2 Hotelling (Jackson, 1991), defined as the Mahalanobis distance between the observation projection onto the principal space and the centre of gravity of the data cloud. T^2 Hotelling, despite Euclidean distance, take into account the covariance matrix between variables (equation 2).

$$T^2 Hotelling_i = \sum_{a=1}^{A} \frac{\mathbf{t}_{iA}^{\,2}}{\mathbf{S}_{tA}^{\,2}} = \begin{bmatrix} \mathbf{t}_{i,1} \\ \mathbf{t}_{i,2} \\ \vdots \\ \mathbf{t}_{i,A} \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{\mathbf{s}_{t1}^{\,2}} & 0 & 0 \\ 0 & \frac{1}{\mathbf{s}_{t2}^{\,2}} & 0 \\ & & \ddots \\ 0 & 0 & \frac{1}{\mathbf{s}_{tA}^{\,2}} \end{bmatrix} \cdot \left[\mathbf{t}_{i,1}, \mathbf{t}_{i,2}, \ldots \mathbf{t}_{i,A}\right] = \mathbf{t}_i^{\,T} \cdot \begin{bmatrix} \frac{1}{\mathbf{s}_{t1}^{\,2}} & 0 & 0 \\ 0 & \frac{1}{\mathbf{s}_{t2}^{\,2}} & 0 \\ & & \ddots \\ 0 & 0 & \frac{1}{\mathbf{s}_{tA}^{\,2}} \end{bmatrix} \cdot \mathbf{t}_i = \mathbf{t}_i^{\,T} \cdot \mathbf{S}_t^{\,-1} \cdot \mathbf{t}_i \quad \textbf{(2)}$$

$\mathbf{t_i}$: S*cores* o observation *i* projection along principal component.

$\mathbf{S_t}$: Covariance matrix.

T^2 Hotelling is distributed with F- Snedecor probability distribution with *A, N-A* degrees of freedom, denoting *A* as the number of principal components and *N* the number of observations. Using this parameter it is possible to establish control limits that will identify severe outliers or observations which behave in an abnormal manner.

Hence, when for a given observation *i*, $T_i^2 > A\cdot(N^2-1)/\, N\cdot(N-A)\cdot F_{\text{critical}}$ (p=0,05), such observation is out of PC control limits with 95% of confidence level. Modifying the confidence level of the PC control chart only requires to recalculate F_{critical} for the appropriate probability value.

- *Moderate outliers* are those whose Euclidean distance to model, or in other words the residual vector module $\|\mathbf{e}_i\|$, is too large (Figure 5b). The statistical parameter used to detect this type of outliers is the Distance to model *(DMOD)* defined by the ratio *Si/So,* where S_i represents the absolute distance to model and S_o the normalized distance of the model.

Therefore, absolute distance and normalized distance of the model are statistics that determine if an observation is too faraway from the model. In such case they will be classified as an outlier.

Absolute distance to model is calculated by:

$$S_i = \sqrt{\frac{\sum_{K=1}^{K} \mathbf{e}_{ik}^{\,2}}{(K-A)}} \quad \textbf{(3)}$$

where

K : Number of primitive variables.

A: Number of Principal Components.

$\mathbf{e}_{ik}$: Observation i residual on variable K.

The normalized distance of the model is an estimation of its residual variability taking into account all the observations used to build the model. It is calculated as follows:

$$S_o = \sqrt{\frac{\sum_{i=1}^{N}\sum_{K=1}^{K}\mathbf{e}_{iK}^{\,2}}{(N-A-A_0)\cdot(K-A)}} \qquad \textbf{(4)}$$

A_0: 1 if model is centered, 0 otherwise.

The absolute distance of one observation divided by the normalized distance to the model squared, $(Si/S_o)^2$, approximate a F-Snedecor probability distribution with *(K-A),(N-A-1)(K-A)* degrees of freedom. This way, the membership probability of one observation can be computed. If $(Si/S_o)^2 > F_{critical}$, then observation *i* can be considered to be out of the control limits for that confidence level. Such observation is suspicious to be an outlier.

Once all outliers are detected and removed from the data matrix, the methodology returns again to the model building step using the debugged data matrix. After several iterations, summarized in Figure 2, a reference PCA-model is calculated.

CASE STUDY: INYECTED FLOW TO AN URBAN WATER UTILITY

INTRODUCTION

This paper exploits the PCA technique to determine hidden correlation structures present in the injected water flow data. This has been done by obtaining a PCA-model that condenses the valuable information contained in the data files stored by the SCADA system of the utility.

As a normal procedure, water companies record all the data coming from the measurement equipment installed in the system. The engineer in charge of controlling network performance supervises the injected flows, specially the injected night flows. Most of the times, the control limits are manually established based on the practical experience of technicians, without using any statistical methods.

Nowadays, the high complexity of water measurement networks, and the number of water district metering areas (DMA) makes this task complicated. Water pressure regulating valves are used to maintain water pressure in between some operational limits, so water may come from different pipes to a given sector. The relations between flow in different pipes change when the opening of these valves is modified. For this reason, it is difficult to handle such quantity of information and upholding coherently water flow control limits to indicate abnormal measures.

To build a model, a stabilization of the process is required to obtain a valid *reference* behaviour. Afterwards, PCA model establishes T^2 Hotelling and DMOD control limits for the measured flows. Flow data which fluctuate in between these limits are considered to respond to a normal behaviour, otherwise, they are considered to be outliers. PCA control charts are appropriate for situations, like in a water supply system, where the process measurements are multivariate and continuous, and where it is important to detect changes in the relationships between variables (Rodriguez and Tobias, 2001). Outlier detection in real-time is one of the most important aims of Multivariate Statistical Process Control (MSPC) (Kourti and MacGregor, 1996).

An important consideration in PCA, and in general in every statistical method, is that *reference* model should be reestimated with a given frequency depending on the network characteristics. Numerous changes in the network operating conditions will imply frequent updates in the PCA-model.

Kurokawa and Bornia (2002a, 2002b) and Harris and Ironmorger (1998) propose new approaches, making use of univariate Statistical Process Control (SPC) to detect water leakage in a distribution network.

Water losses are controlled by a classical univariate statistical analysis of the injected flows. Control limits for common water demands are also established.

In contrast, this paper presents a multivariate SPC methodology to detect failures or unusual water demands in the water network. This methodology is more sensitive to abnormal behaviours, for example leaks, than other univariate statistical methods since it takes into account two parameters: variability and relationships between variables.

PCA-MODEL BUILDING FOR NIGHT WATER FLOWS

In this case study, as already mentioned, the water network SCADA system collects the injected night flows into a given sector of the network. Therefore, the starting point of this analysis consist of selecting the appropriate data and organising it correctly into the raw data matrix.

First of all, to ensure certain homogeneity of the studied data, night flows correspond to a time period in which the operational conditions of the sector were quite homogeneous. During the time interval considered there were no great pressure or water demand changes due to seasonal effects or population increments. The data is organised in rows, corresponding each one to one day. For each day (observation), there are records about the injected water flow from 00:00 a.m. to 6 a.m., in one hour intervals arranged in the following data matrix:

$$\mathbf{Z} = \begin{matrix} Day1 \\ \cdot \\ \cdot \\ DayN \end{matrix} \begin{bmatrix} Q_{1,0} & \cdots & Q_{1,6} \\ \vdots & \ddots & \vdots \\ Q_{N,0} & \cdots & Q_{N,6} \end{bmatrix} \quad \textbf{(5)}$$

where $Q_{i,j}$ means the injected water flow during hour *j* and day *i*.

PCA-model building is carried out iteratively in several rounds, until a *reference* working condition for injected water flows is found. In such case, PCA-model is fitted. Consequently, in this section, the different stages carried out during the model building are briefly described. The statistical computations have been completed using SIMCA-P 9.0 software, although it may be done using a simple home made software.

To begin the analysis, as discussed in the preceding sections, it is decisive to pre-treat the data matrix (Eriksson et al., 1999). Frequently, one of the major problems of this kind of analysis is the heterogeneous characteristics and the substantially different numerical ranges of the variables. Therefore, since PCA is a maximum variance method, a variable with a large variance, i.e. large numerical range, will have a better chance to be modelled as principal component than other variables with less variance. Hence, it is interesting to transform data into a more suitable form of analysis, considering in this case, *autoscaling* procedure -mean centered with each column and scaled to unit- variance - as the most common technique (equation 6).

$$\mathbf{X} = \begin{bmatrix} \dfrac{(Q_{1,0} - \overline{Q_0})}{\sigma_0} & \cdots & \dfrac{(Q_{1,6} - \overline{Q_6})}{\sigma_6} \\ \vdots & \ddots & \vdots \\ \dfrac{(Q_{N,0} - \overline{Q_0})}{\sigma_0} & \cdots & \dfrac{(Q_{N,6} - \overline{Q_6})}{\sigma_6} \end{bmatrix} \quad \textbf{(6)}$$

where $\overline{Q}_{n,k}$ and σ_k are the average night water flow vector and standard deviation vector, respectively.

$$\overline{Q_k} = \frac{\sum_{n=1}^{N} Q_{n,k}}{N} \qquad \sigma_k = \frac{\sum_{n=1}^{N} (Q_{n,k} - \overline{Q_k})^2}{N-1} \quad \textbf{(7)}$$

The next stage, after data matrix pre-treatment, consists of computing the principal components $\mathbf{p_a}$ of the initial raw data matrix. Principal components are obtained by calculating the eigenvectors of the covariance pre-treated data matrix. Its eigenvalues $\lambda_{1,}\lambda_{2,\ldots,}\lambda_A$ are usually listed in decreasing other of importance, $\lambda_1 > \lambda_2 > \ldots > \lambda_A$, with respect to their ability to describe data variability.

The *first principal component* is the direction that best explains the data cloud. In other words, the axis of maximum variation of the points whose projections onto this line minimize their distance to the data in a least squares sense. The s*econd principal component* is the line that passes through the data average points and minimizes the projection distances in a direction that is orthogonal to the first principal component.

Principal components are computed until model adjustment to original data matrix is considered to be sufficient. However, the number of principal components has to be selected carefully. Sometimes, adding a new component to the PCA-model can lead to an over adjustment that deteriorates its predictive capacity. Knowing the number of principal component to include in the PCA-model is a critical stage, called model diagnosis.

For the case studied, each observation, corresponding to the injected night flow, was made up of seven variables. The PCA technique reduced the dimensionality of the variable space from seven to two. *Cross-validation* technique proved that adding a third principal component was not necessary. Therefore, the loading matrix **P** resulted a two column matrix, in which each column is constituted by the corresponding eigenvectors of the covariance matrix as shown in the following expression:

$$\mathbf{P} = \begin{bmatrix} p_0^{\,1} & p_0^{\,2} \\ \vdots & \vdots \\ p_6^{\,1} & p_6^{\,2} \end{bmatrix} \tag{8}$$

Once principal component are identified, it is possible to project the observations on the new two-dimensional space creating the *score* matrix:

$$\mathbf{T} = \mathbf{X_{N,6}} \cdot \mathbf{P_{6,2}} \tag{9}$$

In this situation, first direction of the reference model points out the average night water flow, i.e., the loading vector in this direction has similar weight during all night. Then in practice, those days with high or low $\boldsymbol{t_1}$ are due to abnormal water demand, extremely high or low, and are seen in bidimensional score plot as typical outliers. In contrast, first two hours of the PCA model, 0:00 and 1:00 a.m, have more influence on the second direction than the rest. Consequently, changes in the water demand during these two hours will be rapidly denote.

After the first PCA-model is built, the next step, following Figure 2, is *model validation*. This phase, essentially graphic, identifies those observations out of the control limits which are called outliers. To ensure that the obtained model is not distorted, these observations have to be eliminated from the original data matrix. This debugging process yields to a *reference* PCA-model, undistorted by outliers, which define the *standard* behaviour of water demand during night hours. The outlier detection is completed individually, computing the two statistics presented in this paper, DMOD and T^2 Hotelling.

Figure 6 shows the score plot of the two first PC extracted from the original data matrix and the DMOD plot. Observations which lay out of the T^2 Hotelling control limit are considered *severe outliers*, while points out of DMOD control limit are define as *moderate outliers*.

Both detection methods are complementary. T^2 Hotelling reveals those samples with excessively high or low values of the variables recorded, even in those cases in which the correlation between variables is maintained. To understand this concept consider Figure 6a and 7. Observations 8, 9 and 163 conserve the same shape as the average during the monitored period, i.e. night hours, but water demand is higher.

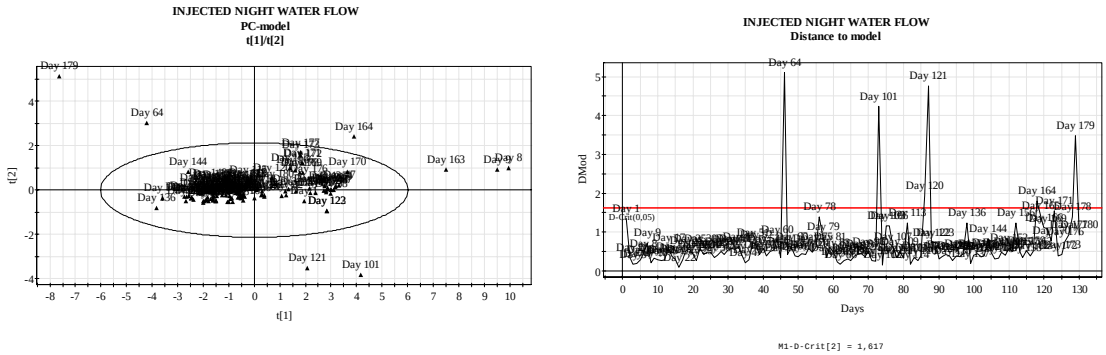

Figure 6. Control chart (a) T^2 Hotelling (b) DMOD .Confidence level 95%.

R^2(cum)=95.2% Q^2(cum)=89.5%

The other parameter used to diagnose the behaviour of an observation is DMOD. This statistic detects sensitively any change in the injected water flow patterns. In other words, it identifies those days where the variable correlation structure is broken. Observations, 64, 121 or 179, depictured in figure 6b and 7, are clear examples. These observations may be classify as both, severe and moderate outlier. The reason for this is that the value of the variables is either too high or too low when compared with the average explained by the first direction. Furthermore, in these particular cases, the correlation structure between variables is also broken.

In Figure 7, all injected flow during the studied period are plotted. Those days with real incidences, with excessive water leakage (pipe bursts) or water service interruptions, effectively exceed the PCA control limits.

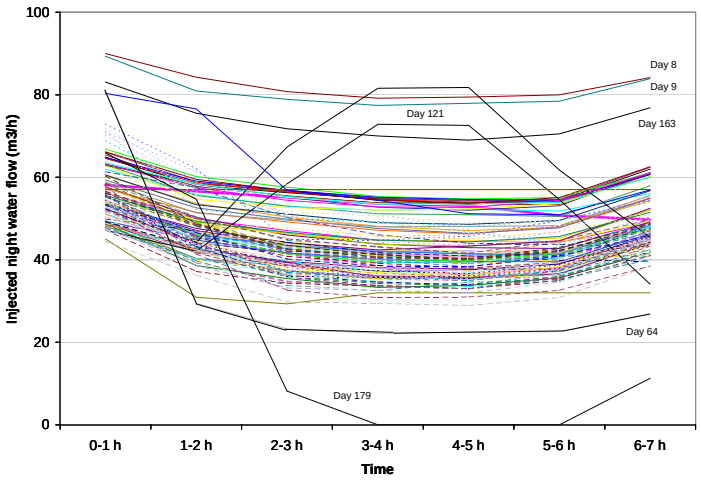

Figure 7. Night water demand pattern of the study period.

An important conclusion extracted from this analysis is that moderate outliers are usually those in which the failure appears between 00:00 a.m. and 6 a.m. Instead, severe outliers, are those observations in which the abnormal water demand started before the studied time interval.

As a result, the final T^2 control chart, as display in Figure 8a, only include the observations used to build the *reference* PCA-model. All of them are inside the T^2 Hotelling limits and scatter homogeneously in the entire ellipse area. This ellipsoid establishes the control limits in a low-dimensional space and defines the allowed variability of the studied process. In contrast, Figure 6a, shows the initial undebugged data cloud concentrated in the centre of the T^2 chart region. In this case control limits are too large due to the presence of outliers in the PCA-model which increase variability.

Nevertheless, in the debugged model it is possible to find some days slightly outside the DMOD control limits (Figure 8b). As a statistical fact, consequence of the 95% confidence level adopted in this case, 5 out of 100 observations may be outside the control limits without being a real moderate outlier.

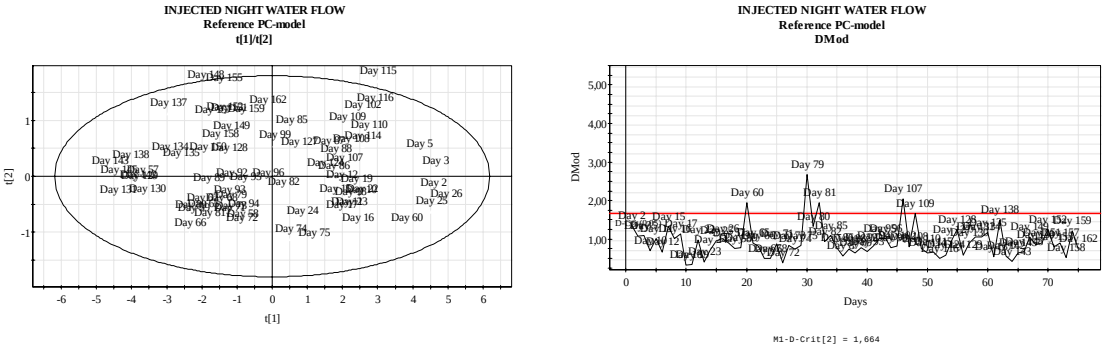

Figure 8. Reference PCA-model control charts.(a) T^2 Hotelling (b) DMOD. Confidence level 95%.

MODEL EXPLOITATION

The use of *reference* PCA-model in the SCADA system is simple. Information about injected water flows is recorded every hour. After the night period these data can be projected on the PCA-model using equation 10. With the calculated projection this observation can be plotted on the T^2 control charts. Likewise, the Euclidean distance to the reference PCA-model can be obtained for this new observation and represented in the DMOD control chart. As describe above, both parameters, T^2 Hotelling and DMOD, are used to evaluate separately each new water flow period to detect possible incidences.

$$\mathbf{t}_{i,a} = \sum_{k=0}^{6} \underset{\sim}{Q}^{injected}{}_{i,k} \cdot p_{k,a} \qquad \textbf{(10)}$$

Figure 9 and 10 depict a simple example that compares the average night water flow of those days included in the reference PCA-model against fictitious day with irregular water demands.

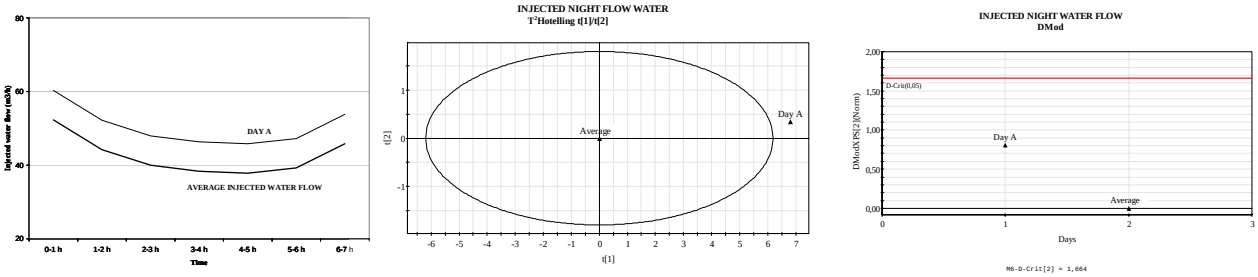

Figure 9. Day A with irregular behaviour. High population water demand .

Figure 9 shows how T^2 Hotelling control chart identifies *day A*, with a large water demand during night hours, as severe outlier. As seen in this figure, for day *A* the shape of water demand curve is similar to the average, i.e. it conserves the variable correlation despite the fact that there is a higher water consumption in this DMA. DMOD parameter is not useful for detecting this type of outliers, distance to model of this observation lays inside control limits.

In contrast, as discussed in preceding sections, *day B* (Figure 10) presents a sudden change in the water demand curve. For this observation the correlation structure between variables is broken, and consequently DMOD for this point is large and out of the defined control limits. T^2 Hotelling control chart does not identify this observation as an outlier.

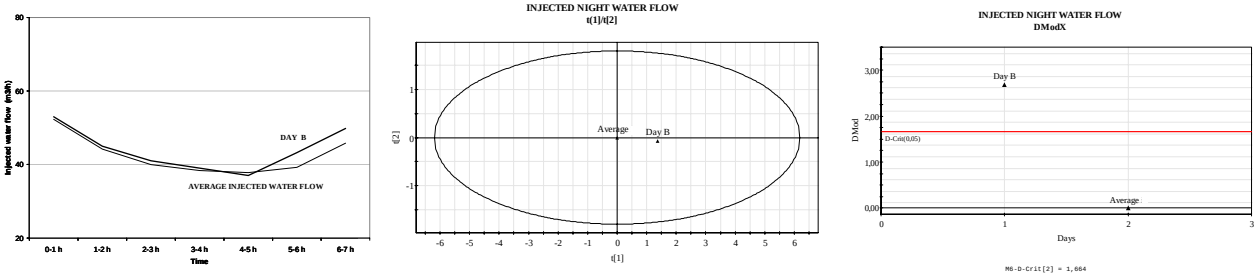

Figure 10. Day B with irregular behaviour. Sudden water pipe burst that produces broken water demand curve.

Another type of control chart, shown in Figure 11, is constructed as a time series of the projected observations along the first principal component, $\mathbf{t}_1$. This chart allows to rapidly identify severe outliers and injected flow tendencies during the entire monitored period. On the one hand, pipe bursts will produce an abnormal high value of $\mathbf{t}_1$. On the other, a communication failure between the flowmeter and the control centre will yield an unusual low value of $\mathbf{t}_1$. Finally, a positive tendency means that leaks in the system are frequent and significantly increase water demand. A steep slope indicates that pipe urgently need to be replaced or rehabilitated.

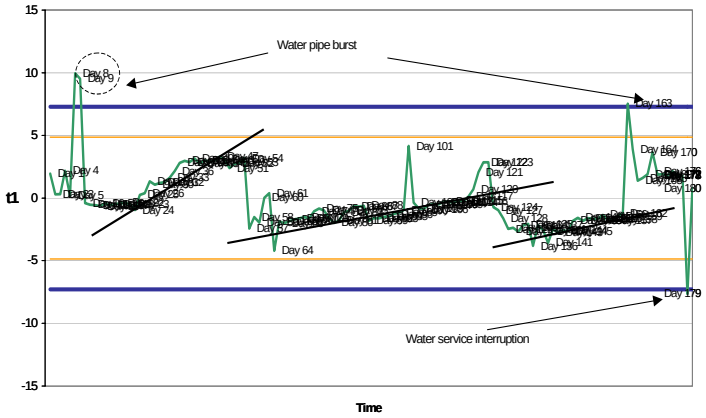

Figure 11. Time-series of projected observations along the first principal component. Confidence level 90 and 95%.

PCA - TECHNIQUE BENEFITS AND CONCLUSIONS

The basis of the methodology presented in this paper is founded in the geometry of the original data cloud, that in this case correspond to the injected water flows during night hours. The PCA-model calculated establishes statistical control limits for T^2 Hotelling and DMOD, which allow to identify irregular behaviours of the system.

The sensitivity of this technique is significantly higher than other univariate statistical methods. The reason is that a univariate approach only takes into account the variability of one variable at a time. In this case, it would mean to solely consider the variability of water consumption during one hour, for example between 5 and 6 a.m. A multivariate method will also consider the relationships between variables, i.e. the correlation structure of night flow water demands. This way, if water consumption during one hour is higher or lower than expected, taking into account measured water demand during previous hours, the multivariate method would detect this irregular behaviour.

Another important benefit, with respect to traditional approaches, is the fact that this statistical procedure is appropriate when process measurements are continuous. Thus, the reaction time of the operator improves

and therefore the economical costs due to water incidences. It is possible to implement different models for several time intervals during the day.

Finally, it is important to mention that this multivariate technique was designed to reduce data complexity. It extracts the most significant information from the numerous initial variables and describes it in a few new variables, converting it into useful and simplified information for the system operators.

Like in any statistical model, PCA-model should be updated periodically to preserve its effectiveness. The update frequency depends on the quantity of changes in the network configuration, the variation in water demands due to seasonal effects or increments in the DMA-water customers, and other local parameters.

As a final conclusion it can be said that multivariate statistical control procedures, and particularly, Principal component analysis, are techniques not frequently used in this field despite their usefulness. PCA contributes very positively in water supply management and can improve operational network conditions since it allows a more rapid and sensitive detection of any incidence produced in the water distribution system.

REFERENCES

ERIKSSON, L.; JOHANSSON, E.;KETTANEH-WOLD, N; WOLD, S. (1999).“Introduction to Multi- and Megavariate data analysis using projection methods”. Umetrics AB.

HARRIS, Ch.; IRONMONGER, R. (1998). “*Socrates-Improving the accuracy of measured night flows”* Driving down water leakage Conference. London.

JACKSON, J.E. (1991). “*A user`s guide to principal components”*. Wiley. New York.

KOURTI, T.; MacGREGOR, J. (1996). “ *Multivariate SPC methods for process and product monitoring*”. Journal of quality technology. Vol. 28, No. 4. pp.409-428.

KUROKAWA, E.; BORNIA, A.C. (2002a). “Utilizando a carta X para avaliaçao de datos diarios da macromediçao de um setor de distribuiçao de agua tratada da cidade de Goiania (Caso Jardim América)”. Seminario Planejamento, projeto e operação de redes de abastecimento de agua. Joao Pessoa. Brasil.

KUROKAWA, E.; BORNIA, A.C. (2002b). “Uma proposta para a utilizaçao do controle estatístico de processo (CEP) a través da carta X como uma ferramenta gerencial para a avaliaçao da vazao mínima noturna de um setor”. Seminario Planejamento, projeto e operação de redes de abastecimento de agua. Joao Pessoa. Brasil.

MALINOWSKI, E.R. (1977). “Determination of the number of factors and the experimental error in the data matrix”. Anal. Chem. 49, 612-617.

MALINOWSKI, E.R. (1987). “Theory of the distribution of error eigenvalues resulting from PCA with applications to spectroscopic data”. J. Chemom. 1 33-40.

RODRIGUEZ, R; TOBIAS, R. (2001). “*Multivariate methods for process knowledge discovery: The power to know your process”*. Statistics, data analysis and data mining. SUGI Proceedings. Paper.252-26.

ROMERO, R. (1997).“Curso de introducción a los métodos de análisis multivariante”. ETSEA. Universidad Politécnica de Valencia

WOLD, H. (1966). NIPALS (Nonlinear Iterative Partial Least Squares). “*Nonlinear estimation by iterative least squares procedures“* , Research papers in Statistics, Wiley, New York, pp 411-444.

WOLD, S. (1978).“Cross validation estimation of the number of components in factor and principal component models”. Technometrics, 20, pp 397-406.

WOLD, S.; ESBENSEN, K.; GELADI, P. (1987).“*Principal Component Analysis*”. Chemometrics and Intelligent Laboratory Systems, 2 , pp. 37-52.

Anonymous (2001). “SIMCA-P 9.0. User guide and tutorial” Ed. Umetrics.