

## **Multi-dimensional visualization and clustering of historical process data**

Nina F. Thornhill\*, Hallgeir Melbø# and Jan Wiik#

\* Department of Electronic and Electrical Engineering, University College London, Torrington Place, London WC1E 7JE (author for correspondence)

# ABB Corporate Research Centre, Bergerveien 12, N-1375 Billingstad, Norway.

### **Abstract**

Multivariate statistical analysis using principal components can reveal patterns and structures within a data set and give insights into process performance and operation. The output medium is usually a two dimensional screen, however, so it is a challenge to visualize the multidimensional structure of a data set by means of a two-dimensional plot.

A method of visualization is described in the form of a hierarchical classification tree that can be used to view the structure within a multivariate principal component model of three or more dimensions. The tree is generated from an unsupervised agglomerative hierarchical clustering algorithm which operates in the score space of the principal component model, and a recursive algorithm to draw the tree. It is readily adaptable to a wide range of multivariate analysis applications including process performance analysis and process or equipment auditing. Its application are illustrated with industrial data sets.

Keywords: Clustering; dendrogram; fault detection; hierarchical classification; performance analysis; plant-wide; principal component analysis; process operation; process monitoring; statistical process control.

## **1 Introduction and motivation**

Large data bases are being accumulated by companies operating oil, gas and chemical processes. These data bases are packed with process measurements and, increasingly, with other measurements such as those from monitoring of rotating machinery, records of energy usage and emissions. Generally the measurements are sampled over time and the raw data would be presented in the form of time histories or sampled data sequences.

Many of the measurements have correlations [1, 2] which can be captured during normal operation and exploited in the detection of abnormal situations. This is the motivation behind the many applications of multivariate analysis to process data [3, 4, 5, 6]. As the data sets become larger and larger, however, it becomes more challenging to present the results of a multivariate analysis. A large integrated process such as an oil platform may have several different modes of behaviour, each of which is significant. While principal component analysis (PCA) might reduce several hundred measurements to, say, ten principal components there then remains an issue of presenting the ten-dimensional model to the analyst so that groups of measurements or episodes of operation having similar characteristics can be isolated and examined in detail. This paper describes a solution to the problem of visualization of the clusters in a high-dimensional PCA model by means of a hierarchical classification tree showing the structure of the PCA score space and which provides a simple and automated way to display the presence of clusters. The key elements in the procedure are the generation of a suitable distance measure, an agglomerative hierarchical clustering algorithm and a recursive algorithm to draw and analyze the clusters within the tree.

The visualization of high dimension multivariate PCA models has previously been examined by Wang *et. al.*, [7] who used parallel coordinates to display multiple dimensions of the score space. They identified abnormal days of running by manual inspection of outliers compared to the bulk of the parallel coordinates plot. A distinctive feature of the work presented here is that it automatically reveals the detailed structure of clusters within the model and gives automated detection of the outliers. The resulting visualization is illuminating and easily understandable.

The next section of the paper gives a review of related work to place the methods in context. Section 3 presents the multivariate data analysis where a distinction is made between the formulations of PCA for analysis of process performance where the operating profiles at each minute, hour or day are the items of interest, and PCA for auditing of process operation in which the time histories of the measurements from each sensor are the items of interest. Distance measures for clustering analysis are also discussed in Section 3 together with the automated algorithm for creation of the hierarchical classification tree and some refinements for easing the handling of very large data sets. Two industrial data sets are then analyzed to illustrate the concepts.

## 2 Background and context

### 2.1 Principal component analysis

Descriptions of principal component analysis may be found from many sources, for example in [8] and [9]. In analytical chemistry, near infrared (NIR) and nuclear magnetic resonance (NMR) spectroscopy data are routinely analysed by PCA for estimation of analyte concentrations in unknown samples [10, 11] and Seasholtz [12] described the industrial application of multivariate calibration in NIR and NMR spectroscopy at Dow Chemical Company. Principal component analysis has proved useful in other diverse areas such as paint colour analysis [13] and in the analysis of the relationship between the crispness of apples and recorded chewing sounds [14]. Applications of principal component analysis are also well established in water quality analysis [15]. Examples include the discovery of temporal and spatial variations in water quality and the pinpointing of sources of river pollution [16], while [17] commented on the value of PCA in condensing and interpreting large amounts of water quality data, finding that the principal components related underlying mechanisms such as biological activity and seasonal effects. Bioinformatics is another area where PCA is having an impact [18].

Industrial uses include principal component analysis for monitoring of machinery and process equipment. Wu *et. al.* [19] used a method known as “eigenfaces” in the recognition of sounds from car engines. PCA has also been used to identify and classify the severity level of bearing defects in rotating machinery [20] and for acoustic monitoring of processes [21]. The classification of the spectra of acoustic signals using multivariate statistical analysis was described in a patent [22] which gave examples of the classification of the acoustic spectra from a pump and an industrial blender.

All the above applications have the common aim to discover structure within the data sets, to ascertain the items within the data sets that belong together and to relate the results to underlying mechanisms.

Applications to process monitoring have been reported extensively [1-6, 23, 24]. A specialized area in process monitoring is on-line multivariate statistical process control in which new measurements are projected into a PCA calibration model that was developed during normal operation. Multivariable warning and alarm limits are set which test whether a new set of measurements is within the normal bounds captured by the calibration model [4, 25, 26].

The work in this paper, like references [10-22], concerns the discovery of structures in the data sets and ascertains the items that belong together. It achieves multidimensional visualization of a complete data set and gains insights by exploration of the structures within the data set. Applications for the approach are suggested in Section 4 where case studies are presented.

## 2.2 Pattern classification and visualization

*Pattern classification:* The book by Duda, Hart and Stork [27] gives a comprehensive review of methods for finding patterns and structure within a multivariate data set. The main division is between supervised and unsupervised methods. The former use a training set of items that have already been categorized by drawing on some prior knowledge, while in the unsupervised methods none of the categories of the items is known in advance. The categorization step comes last in an unsupervised method; the analysis draws attention to the groupings and structure in the data set and the analyst then inspects and categorizes the groups. However, even unsupervised methods work best if some *a-priori* knowledge is exploited such as an idea of what constitutes similarity of the items.

Examples of unsupervised clustering methods are agglomerative hierarchical clustering, partitioning methods such as *k*-means clustering, and density based methods. PCA and other multivariate statistical methods were highlighted alongside unsupervised clustering methods by Oja [28] as having the capacity to give insights into the true nature and structure of the data. This paper uses agglomerative hierarchical clustering within the score space of a principal component analysis to detect the structure within a data set.

*Visualization using hierarchical classification trees:* Gordon [29] gave a comprehensive review of hierarchical classification, distinguishing between agglomerative and divisive methods. Agglomerative hierarchical clustering is an unsupervised algorithm for building up groups of similar items from a population of individual items. The basic algorithm [27] starts with *N* clusters each containing one item and proceeds as follows:

```
repeat
    find the pair of nearest clusters
    merge them into one cluster
until there is one cluster containing N items
```

The results of agglomerative hierarchical clustering may be visualized in a classification tree in which the items of interest are the leaves on the tree and are joined into the main tree and eventually to the root of the tree by branches. A classification tree shows the structure while a dendrogram has branches of various lengths to indicate the degree of similarity between items and sub-clusters. Classification trees are also used in divisive classification in which a large group of items is recursively split into subcategories.

Industrial applications of clustering and/or classification trees have included methods for office buildings to detect days of the week with similar profiles of energy use [30], the presentation of results from an end-point detection method in a crystallization process [31] and from analysis of illegal adulteration of gasoline with organic solvents [32]. In the area of process analysis, hierarchical classification has been combined with PCA for detection of key factors that affect process performance in a blast furnace and a hot stove system generating hot air for the blast furnace [33, 34]. A divisive classification was used in which clusters of items appearing in the score space of the first two principal components were identified and then further divided into sub-clusters using PCA on the identified clusters.

The distinctiveness of the method presented in this paper is that it uses agglomerative classification in the score space of all significant principal components, and it is fully automatic. It takes multivariate analysis to a new level in process monitoring to achieve visualization and detection of high-dimensional PCA clusters.

### 3 Mathematical formulations and distance measures

#### 3.1 Multivariate methods

*PCA for process performance analysis:* In multivariate process performance analysis each row of the  $N \times m$  data matrix,  $\mathbf{X}$ , is a sample of  $m$  measurements. There are  $N$  such samples where in general  $N > m$ . The columns of  $\mathbf{X}$  are mean centred and scaled to unit standard deviation. Each row of  $\mathbf{X}$  is a *plant profile*, a snapshot at a particular time of the pattern of measurements across the plant. Of the numerical values in each row of  $\mathbf{X}$ , any positive values indicate sensors reading above average at that time and negative values are sensors reading below average. The plant profiles do not in general have zero mean or unit standard deviation, for instance it is possible for all sensors to be reading above average at a particular time.

$$\mathbf{X} = \begin{matrix} m \text{ measurements} & \rightarrow \\ \begin{pmatrix} x_1(t_1) & \dots & x_m(t_1) \\ \dots & \dots & \dots \\ x_1(t_N) & \dots & x_m(t_N) \end{pmatrix} & \begin{matrix} N \text{ time} \\ \text{samples} \\ \downarrow \end{matrix} \end{matrix} \quad (1)$$

The singular value decomposition is  $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$  where matrix  $\mathbf{U}$  has orthogonal columns and  $\mathbf{V}^T$  has orthogonal rows,  $\mathbf{U}$  is  $N$ -by- $m$ ,  $\mathbf{V}^T$  is  $m$ -by- $m$ , and  $\mathbf{D}$  is diagonal and its elements are the positive square roots of the eigenvalues of the  $m$ -by- $m$  matrix  $\mathbf{X}^T\mathbf{X}$ . The principal component decomposition is  $\mathbf{X} = \mathbf{T}_m\mathbf{W}_m^T$ , where  $\mathbf{T}_m = \mathbf{U}\mathbf{D}$  and  $\mathbf{W}_m^T = \mathbf{V}^T$ . The subscript  $m$  indicates a full rank PCA model in which all  $m$  principal components are included.

A description of the majority of the variation in  $\mathbf{X}$  can often be achieved by truncating the PCA description. A  $p$ -principal component model is  $\mathbf{X} = \mathbf{T}_p\mathbf{W}_p^T + \mathbf{E}$  in which the variation of  $\mathbf{X}$  that is not captured by the first  $p$  principal components appears in an error matrix  $\mathbf{E}$ . The  $\mathbf{w}'$ -vectors are rows<sup>1</sup> of  $\mathbf{W}_m^T$ .

$$\begin{aligned} \mathbf{X} &= \begin{pmatrix} t_{1,1} \\ \dots \\ t_{N,1} \end{pmatrix} \begin{pmatrix} w_{1,1} & \dots & w_{1,m} \end{pmatrix} + \begin{pmatrix} t_{1,2} \\ \dots \\ t_{N,2} \end{pmatrix} \begin{pmatrix} w_{2,1} & \dots & w_{2,m} \end{pmatrix} + \dots + \begin{pmatrix} t_{1,p} \\ \dots \\ t_{N,p} \end{pmatrix} \begin{pmatrix} w_{p,1} & \dots & w_{p,m} \end{pmatrix} + \mathbf{E} \\ &= \begin{pmatrix} t_{1,1} \\ \dots \\ t_{N,1} \end{pmatrix} \mathbf{w}'_1 + \begin{pmatrix} t_{1,2} \\ \dots \\ t_{N,2} \end{pmatrix} \mathbf{w}'_2 + \dots + \begin{pmatrix} t_{1,p} \\ \dots \\ t_{N,p} \end{pmatrix} \mathbf{w}'_p + \mathbf{E} = \mathbf{T}_p\mathbf{W}_p^T + \mathbf{E} \end{aligned} \quad (2)$$

---

<sup>1</sup> The vector notation in use in this paper is that  $\mathbf{z}$  represents a column vector and  $\mathbf{z}'$  indicates a row vector.

The model uses  $p$  rows of  $\mathbf{W}_m^T$  each with  $m$  elements which are called *loadings*. They act as a set of orthonormal basis functions from which all the plant profiles (the rows of the  $\mathbf{X}$  matrix) can be approximately reconstructed. The  $\mathbf{t}$  – vectors are column vectors each containing  $N$  elements which are called *scores*. The scores indicate the amplitude of each normalized basis function in the reconstruction. The aim of PCA is to choose a reduced number of terms,  $p$ , so that the reconstruction shown in equation (2) adequately captures the main patterns in the plant profiles in the data set while rejecting noise. The choice of the number of terms in this paper has been made using the average eigenvalue method which is simple to implement and worked well in comparisons made by Valle *et al.* [35].

Clusters in process performance analysis: If the process has the same operating mode at two different sampling times then the scores tend to be similar to one another such that, for instance, vectors  $\mathbf{t}'_i = (t_{i,1}, t_{i,2}, \dots, t_{i,p})$  and  $\mathbf{t}'_j = (t_{j,1}, t_{j,2}, \dots, t_{j,p})$ , the  $i^{\text{th}}$  and  $j^{\text{th}}$  rows of the  $\mathbf{T}_p$  matrix, are similar to each other if the plant profile at sample time  $i$  is similar to that at sample time  $j$ . There may also be other similar rows in the  $\mathbf{T}_p$  matrix if the plant profile described by  $\mathbf{t}'_i$  and  $\mathbf{t}'_j$  is a common one, hence clusters and structure in the data set can be detected from similarity of the scores. The visualization issue in PCA for multivariate monitoring is to observe these clusters of plant profiles. A common way of visualization is by means of two dimensional plots of one score versus another. Each spot in this two-dimensional score plot represents the projection of a plant profile at a particular instant of time onto the two dimensional score space. Two scores are not usually enough to capture all the important features of the plant profile, however, which is the motivation for the use of a classification tree. Reference [7] presented the  $\mathbf{t}'$  – vectors on a set of parallel axes which enabled all scores values to be seen simultaneously, but then each  $\mathbf{t}'$  – vector is represented by a series of line segments rather than a spot and it is difficult to see the structure within the data set.

PCA for process audit: In a process audit the items of interest are the time histories of the measurements (the measurements are also referred to as *tags*). The basis functions for the time histories are the columns of matrix  $\mathbf{U}$  in the singular value decomposition, a technique known as classical scaling [8]. It is convenient in practice to be able to use the same PCA codes as for process performance analysis, which can be achieved by transposing the data matrix. The rows of  $\mathbf{X}$  are now the time histories. The rows (not the columns) of  $\mathbf{X}$  are scaled to zero mean and unit standard deviation, thus the data matrix  $\mathbf{X}$  has different properties in a process audit compared to a plant performance analysis.

$$\begin{array}{c}
 N \text{ time samples} \rightarrow \\
 \mathbf{X} = \begin{pmatrix} x_1(t_1) & \dots & x_1(t_N) \\ \dots & \dots & \dots \\ x_m(t_1) & \dots & x_m(t_N) \end{pmatrix} \begin{array}{c} m \\ \text{measurements} \\ \downarrow \end{array} \quad (3)
 \end{array}$$

The  $\mathbf{w}'$  – basis vectors in (2) now have  $N$  elements and resemble time histories and each time history in a row of  $\mathbf{X}$  is built up from a linear combination of these basis functions. Each row of  $\mathbf{X}$  maps to

a spot in the score plot and clusters in the scores may then be used to detect clusters of tags whose time histories have similar characteristics.

### 3.2 Distance measures

*Euclidian and angle measures:* The Euclidian distances between scores in the score plot has been used as a measure for detection of PCA clusters. If  $p$  principal components were being used then the co-ordinates of the vector  $\mathbf{t}'_i$  joining the origin to the  $i$ 'th spot in the score plot are:

$$\mathbf{t}'_i = (t_{i,1}, t_{i,2}, \dots, t_{i,p}) \quad (4)$$

The Euclidian distance between  $\mathbf{t}'_i$  and  $\mathbf{t}'_j$  is:

$$d_{i,j} = \sqrt{\sum_{k=1}^p (t_{i,k} - t_{j,k})^2} \quad (5)$$

Some sets of process data form tight clusters with small Euclidian distances. This typically happens when a process moves between distinct operating states.

In process performance analysis the angular measure discussed in [27] is often more suitable than the Euclidian distance. The reason for this observation is that the PCA clusters frequently take the form of plumes radiating from the origin. Raich and Cinar [36] observed plumes in their analysis of simulated faults in the Tennessee Eastman benchmark model. An example, to be discussed in more detail later, can be seen in the two dimensional score plot of Figure 4 where days 323, 90, and 106 appear in a line along the Score 3 ( $t_3$ ) axis. In fact, day 129 also appears to be in the same plume but, as discussed later, day 129 becomes separated when more principal components are included.

The physical meaning and reason for the plume is that the plant profiles on these days resemble the third loading vector, as can be seen in the original data on close inspection (the profiles for days 323, 90, and 106 are presented in Figure 2 and the third loading vector is the third line from the top in Figure 3). The smaller  $t_3$  score for Day 106 in Figure 4 means that the profile was present on that day but not as intensely as on day 323 when the  $t_3$  score was larger.

Both [36] and [37] used the sum of the angles between the principal components of calibrated historical data sets and data from recent operation. The aim was to match the recent plant profiles with a historical data set from a known fault. Here, by contrast, the angles between individual items within one data set are used in order to reveal the multi-dimensional structure.

*Formulation of an angular measure:* An appropriate measure for membership of a plume is that the direction of vector  $\mathbf{t}'_i$  in the multidimensional score plot lies within the same solid angle as those of other  $\mathbf{t}'$ - vectors belonging to items within the plume. The angle between  $\mathbf{t}'_i$  and  $\mathbf{t}'_j$  may be determined through calculation of the scalar product:

$$\cos(\theta_{i,j}) = \frac{\mathbf{t}'_i \cdot \mathbf{t}'_j}{\|\mathbf{t}'_i\| \|\mathbf{t}'_j\|} \quad (6)$$

where:

$$\mathbf{t}'_i \cdot \mathbf{t}'_j = \sum_{k=1}^p t_{i,k} t_{j,k} \quad \text{and} \quad |\mathbf{t}'_i| = \sqrt{\sum_{k=1}^p t_{i,k}^2}$$

The vector  $\mathbf{t}'_i = (t_{i,1}, t_{i,2}, \dots, t_{i,p})$  is the  $i$ 'th row of matrix  $\mathbf{T}_p$  in  $\mathbf{X} = \mathbf{T}_p \mathbf{W}_p^T + \mathbf{E}$  when a model with  $p$  principal components is in use. Therefore the following calculation steps lead to a symmetric matrix whose elements are the wanted angles  $\theta_{i,j}$ :

**Algorithm: Calculation of the angle measures**

Step 1: Create a normalized matrix  $\tilde{\mathbf{T}}$  from  $\mathbf{T}_p$  whose row vectors are of unit length. Each element in the

$i$ 'th row of  $\mathbf{T}_p$  is divided by  $\sqrt{\sum_{k=1}^p t_{i,k}^2}$  where  $p$  is the number of principal components in use.

Step 2: Determine the matrix  $\mathbf{C} = \tilde{\mathbf{T}} \tilde{\mathbf{T}}^T$ . The elements of  $\mathbf{C}$  are  $\cos(\theta_{i,j})$ .

Step 3: Create the matrix of angles from  $\mathbf{A} = \arccos(\mathbf{C})$ , which returns results in the range 0 to 180 degrees.

Relationship to the correlation coefficient: If the rows of data matrix  $\mathbf{X}$  have been normalized, as is the case in the process audit formulation of PCA, then matrix  $\mathbf{C}$  is identical to the matrix of correlation coefficients between the rows of the reconstructed data matrix  $\mathbf{X}_p = \mathbf{T}_p \mathbf{W}_p^T$ . The correlation coefficient between rows  $\mathbf{x}'_{p_i}$  and  $\mathbf{x}'_{p_j}$  is defined as

$$r_{i,j} = \frac{(\mathbf{x}_{p_i} - \bar{\mathbf{x}}_{p_i}) \cdot (\mathbf{x}_{p_j} - \bar{\mathbf{x}}_{p_j})}{|\mathbf{x}_{p_i}| |\mathbf{x}_{p_j}|}$$

where  $\bar{\mathbf{x}}_{p_i}$  is the mean value of the elements in the  $i$ 'th row of  $\mathbf{X}_p$ . In the process audit formulation the rows of  $\mathbf{X}_p$  have zero mean because the  $\mathbf{w}'$ -basis functions have zero mean and the correlation coefficient is therefore given by:

$$r_{i,j} = \frac{\mathbf{x}_{p_i} \cdot \mathbf{x}_{p_j}}{|\mathbf{x}_{p_i}| |\mathbf{x}_{p_j}|} \quad (7)$$

Substitution of  $\mathbf{x}_{p_i} = t_{i,1} \mathbf{w}'_1 + t_{i,2} \mathbf{w}'_2 + \dots + t_{i,p} \mathbf{w}'_p$  and exploiting the orthogonality of the  $\mathbf{w}'$ -vectors leads to:

$$\begin{aligned} r_{i,j} &= \frac{(t_{i,1} \mathbf{w}'_1 + t_{i,2} \mathbf{w}'_2 + \dots + t_{i,p} \mathbf{w}'_p) \cdot (t_{j,1} \mathbf{w}'_1 + t_{j,2} \mathbf{w}'_2 + \dots + t_{j,p} \mathbf{w}'_p)}{\sqrt{\sum_{k=1}^p t_{i,k}^2} \sqrt{\sum_{k=1}^p t_{j,k}^2}} \\ &= \frac{\sum_{k=1}^p t_{i,k} t_{j,k}}{|\mathbf{t}'_i| |\mathbf{t}'_j|} = \frac{\mathbf{t}'_i \cdot \mathbf{t}'_j}{|\mathbf{t}'_i| |\mathbf{t}'_j|} \end{aligned} \quad (8)$$

This insight into the statistical nature of the angle measure will be used in the case study in section 4. In the process performance application the rows of  $\mathbf{X}_p$  are reconstructed plant profiles which do not generally have zero mean and the above comments do not apply.

### 3.3 Clustering

*Automatic identification of PCA clusters:* The matrix  $\mathbf{A}$ , whose elements are  $\theta_{i,j}$ , is to be analyzed to find high-dimensional plumes in the PCA score plot. Two items in the score plot whose  $\mathbf{t}'$ -vectors point in similar directions give a small value of  $\theta_{i,j}$ . A fully automated agglomerative hierarchical clustering algorithm based on [8] is used:

**Algorithm: Agglomerative classification**

Step 1: The starting point is the matrix of angular distances with elements  $\theta_{i,j}$ . A text vector of row and column headings is also defined which initially is (1 2 3 4 5 ...) to keep track of the items in the data set. For an process performance analysis application the items are the  $N$  plant profiles in the data set, for a process audit the items are the  $m$  tags.

Step 2: At the  $k$ 'th iteration, the smallest non-zero value  $\theta_{i,j}$  in the matrix is identified. Its row and column indexes  $i$  and  $j$  indicate the smallest angular separation and these are clustered together.

Step 3: A smaller matrix  $\mathbf{A}_k$  is then generated from the original. It does not have rows and columns for the two similar items identified at step 2. Instead, it has one row and column that give the distances of all the other items from the cluster. The distances are  $\min\{\theta_{i,n}, \theta_{j,n}\}$ , i.e. the angular distance between the  $n$ 'th item and whichever member of the cluster was closer. For instance, if  $\theta_{9,15}$  is the smallest angular separation in the matrix then rows 9 and 15 would be deleted and replaced by a new single row, and likewise for columns 9 and 15.

Step 4: The row and column headings are redefined. The heading for the new row created at step 3 indicates the items that have been combined. For instance, if the smallest angular separation at Step 3 had been  $\theta_{9,15}$  then the heading of the new row and column would be (9 15).

Step 5: The results of the  $k$ 'th step are written to a report showing the cluster size, the row heading for the cluster formed at iteration  $k$ , and the two sub-clusters within it. An example is presented in the appendix.

Step 6: Steps 2 to 5 are repeated until all the items have been clustered. At any stage, the outcome of the next step is either another item added to a cluster already identified or the combining of two items to start a new cluster.

A feature of the agglomerative hierarchical classification procedure presented here is that it provides a text-based report which enables the detection of significant clusters as well as automated generation of the hierarchical tree plot.

*Plotting of the hierarchical tree:* The graphical representation of the hierarchical tree can be extracted from the report generated by the above algorithm. The tree is a dendrogram because it represents the sizes of the clusters on the vertical axis. It utilises an algorithm which starts at the top and systematically searches down the left and then the right branches and sub-branches to parse the structure of the tree. The algorithm is recursive meaning it calls itself over and over again in a nested way until it reaches a leaf of the tree. The end result is a set of  $x$ - and  $y$ - coordinates tracing the path that joins each individual item on the horizontal axis to the master node at the top of the tree, which are then plotted using a staircase plot. The number of steps is dominated by the square of the number of items in the tree.



Some nodes in the classification tree have more than two sub-branches, for example Days 138, 140 and the (139, 142) subcluster in the upper panel of Figure 5 are joined by a single horizontal line. The reason is that the overall maximum distance between items in the cluster does not always grow when a new item is added. There is a brief worked example in the Appendix.

Automatic detection of significant clusters: The hierarchical tree is full of clusters because branches join together at each iteration of the algorithm, so it is necessary to detect the significant clusters which reveal the most information about the application. The criterion used to define a significant cluster is the ratio between the length of the branch connecting a cluster of items into the main tree and the maximum distance between the items in the cluster (both measured on the y-axis scale). A ratio of 1 is a good general-purpose threshold. Although this is subjective, the aim is to capture what an analyst might decide in a visual inspection of the tree. For instance, days (9 15 29 22) in Figure 8 are joined to each other low down on the vertical axis scale giving a tight group with small distances between the items. They are connected by a long branch to the rest of the tree so the branch ratio is greater than 1 and therefore they form a significant cluster.

The automated algorithm for generation of the hierarchical tree reports the y-axis values for the horizontal lines joining clusters together and into the main tree. The detection of significant clusters therefore involves the straightforward calculation of the branch length by subtraction and the comparison with the size of the cluster. An example is given in the appendix.

Useful features of the tree are that data sets with no strong structure do not yield significant clusters and that not all data points have to be a member of a cluster. Items that are unique in the data set typically lie by themselves with no close neighbours and are joined into the main tree by long branches.

Refinement for large data sets: A process performance analysis application with many days of running can generate many hundreds of plant profiles, and a plant audit of a very large process or site may have many hundreds of tags. As indicated by Yang *et. al.* [38], the best strategy is to put fewer items into the plot. In their work they presented several multivariable visualizations methods including hierarchical scatter plots and hierarchical parallel coordinates in which the data are first clustered and then the clusters become the items in the plot. Additional plots are then needed, however, to present the structure within the cluster. The benefit of the hierarchical classification tree is that it shows all the structure in one plot, but pruning is necessary to select the relevant items.

One way of pruning the tree is to use only the longest  $\mathbf{t}'$ -vectors in the tree (equation (4)). This method is useful when the clusters lie in plumes and an angular distance measure is in use, it means that the clusters are created using the items further out towards the ends of the plumes. The length of

the  $i^{\text{th}}$   $\mathbf{t}'$ -vector in an analysis with  $p$  principal components is  $|\mathbf{t}'_i| = \sqrt{\sum_{k=1}^p t_{i,k}^2}$ . They are sorted by length and the longest ones are selected to go into the clustering algorithm. Once clusters involving the longest  $\mathbf{t}'$ -vectors have been identified then the  $\mathbf{t}'$ -vectors for the other items can be examined to find any whose angles lie within an already-identified significant cluster.

Other cases call for a different approach. A process which moves between distinct operating states was mentioned earlier. In that case the appropriate way of reducing the number of items is to use fewer plant profiles. The data set may typically capture 75 plant profiles from one operating state even though many fewer, say five, are enough to establish the operating state as a distinct cluster. The number of items can then be reduced by using every fifteenth row of the data matrix  $\mathbf{X}$ . The

hierarchical tree from the reduced matrix will indicate the timings of the transitions between operating states which can then be explored in more detail if required.

*Computation speed:* The speed of computation for PCA and for cluster reporting is high and data matrices with rank of more than 500 have been handled comfortably on a 1.2GHz processor. The need for data reduction discussed above is for the benefit of plotting and visualization of the hierarchical tree. Too many items make the tree too dense and also increases the time to run the recursive algorithm for plotting the tree.

### 3.4 Outlier detection

*Automatic identification of outliers in process performance analysis:* In process performance analysis, all the data including days with abnormal operation are analysed together in one data set and the PCA model therefore captures the significant features of the plant profiles of both the normal and the abnormal days. There is no calibration model representing normal process operation with which abnormal days are compared. Therefore the SPE and Hotelling  $T^2$  measures [4, 25, 26] which are useful in on-line multivariate statistical process control are not appropriate for the detection of the abnormal days in the application presented here. The SPE for the plant profile of the  $i^{\text{th}}$  day is  $|\mathbf{e}'_i|^2$ , where  $\mathbf{e}'_i$  is the  $i^{\text{th}}$  row of  $\mathbf{E}$  in (2). However,  $|\mathbf{e}'_i|^2$  should always be small because the PCA model captures the abnormal as well as the normal days. The Hotelling  $T^2$  measure is not appropriate either because no assumptions can be made about its statistical distribution when abnormal data are included in the  $\mathbf{X}$  matrix.

A non-parametric method which makes no statistical assumptions is more appropriate. The following method uses a percentile limit. The definition of an abnormal plant profile is one that has an extreme value outside of the 5<sup>th</sup> and 95<sup>th</sup> percentiles in any one of its scores. The percentile thresholds are adjustable and the percentage of abnormal items detected depends on the percentiles selected. Choosing the 10<sup>th</sup> and 90<sup>th</sup> percentiles would result in more days classified as abnormal, choosing the 1<sup>st</sup> and 99<sup>th</sup> would result in fewer. The inspiration for this automated algorithm came from the manual procedure applied in [7].

#### **Algorithm: Automated detection of outliers**

Step 1: Each column  $\mathbf{t}_j$  of the  $N \times p$  score matrix  $\mathbf{T}_p$  is sorted into increasing order and examined to find the 5<sup>th</sup> and 95<sup>th</sup> percentiles. The range of score values between these two limits is calculated as  $s_j = t_{j,95\%} - t_{j,5\%}$  where  $t_{j,95\%}$  is the score at the 95<sup>th</sup> percentile and  $t_{j,5\%}$  the score at the 5<sup>th</sup> percentile.

Step 2: A new scaled matrix  $\tilde{\mathbf{T}}$  is created whose  $i^{\text{th}}$  column is  $\mathbf{t}_i/s_i$ . Most of the entries in this matrix lie between  $\pm 1$ , and any entries beyond  $\pm 1$  are outliers. The purpose of the scaling procedure is to identify outliers within each column of  $\mathbf{T}_p$  and also to make it possible to compare outliers in different columns of the score matrix.

Step 3: Any entries in the scaled matrix  $\tilde{\mathbf{T}}$  having absolute values greater than 1 are classified as belonging to abnormal plant profiles. For each such entry, the row in which it is found indicates the row of the abnormal plant profile in the data matrix  $\mathbf{X}$ .

Step 4: An *outlier index* is determined for each plant profile as the largest absolute value in the corresponding row of the  $\tilde{\mathbf{T}}$  matrix. For instance if  $t_{12,5} = 1.1$  and  $t_{12,3} = -1.6$  it means day 12 has an outlier index of 1.6 and the outlier is in score 3.

A possible alternative is to treat the items with the longest  $\mathbf{t}'$ -vectors as outliers in the same way as the longest  $\mathbf{t}'$ -vectors are used to aid visualization for a large data set. It is not the case, however, that a large  $\mathbf{t}'$ -vector means abnormal operation. The argument is a geometric one based on the observation that the diagonal of a hypercube is longer than the sides of the hypercube. The hypotenuse of a right-angled equilateral triangle is longer by a factor of  $\sqrt{2}$  than its sides, and in four dimensions the factor is 2. The  $\mathbf{t}'$ -vector of an item in the data set with mixed behaviour having moderate scores for several different basis functions in equation (2) can therefore become quite long without having any extreme behaviour.

*Outliers in process audit:* The notion of an outlier does not apply in a process audit. If one tag has an extreme value in one score (say, the  $i$ 'th) it should not be interpreted as abnormal. Rather the interpretation is that this measurement is detecting something unique whose time trend is captured in the  $i$ 'th  $\mathbf{w}'$ -basis vector.

## 4 Case study 1 - wastewater

### 4.1 The data set and motivations for analysis

Figure 1 shows the Barcelona wastewater benchmark data set from the University of California at Irvine [39] which has previously been analyzed in the literature [7, 39, 40]. After exclusion of days with incomplete data (147 of them), there are complete data records for 380 days of operation giving the daily averages of 38 measured variables. As pointed out in [7], the daily averages are steady state data and there is no dynamical component present, moreover the data are not available every day and the horizontal day-number axis in Figure 1 is not a continuous time axis. Several days of abnormal operation can be viewed in the data set, for instance days 33, 63 and 153 show large deviations.

Figure 2 shows a selection of the plant profiles. In this figure, the horizontal axis represents the measured variables. The piecewise linear profiles show which measurements were above and which below the average on a particular day. For instance, on day 350, the measurement from Tag 2 was greatly above average while those from Tags 16 and 23 were below average.

The case study first represents the structure within the score space as a hierarchical tree to show clusters of days with similar operation, and then carries out automated detection of abnormal days. The benefit of such an analysis is that it gives insights into the *performance* of the process, including detection of days within the normal range but which had similarities to an abnormal day. This information might be compared with weather reports, local events such as pollution incidents, or faults in the process to determine the range of shocks the process can tolerate before performance is classed as abnormal.

A plant audit has also been completed after the abnormal days have been removed to show clusters of tags which move together in a coordinated manner. The audit provides insights into the *operation* of the process, reveals measurements connected by underlying physical and chemical mechanisms and suggests where measurement redundancy exists.

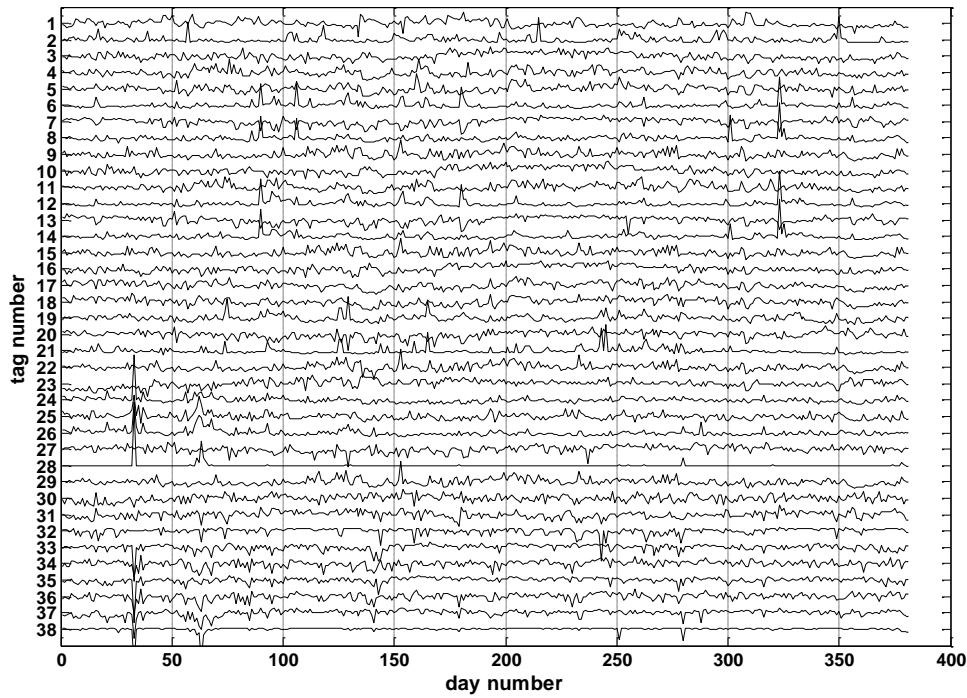


Figure 1. Daily averages from days with no missing data of 38 measurements from the benchmark wastewater data set.

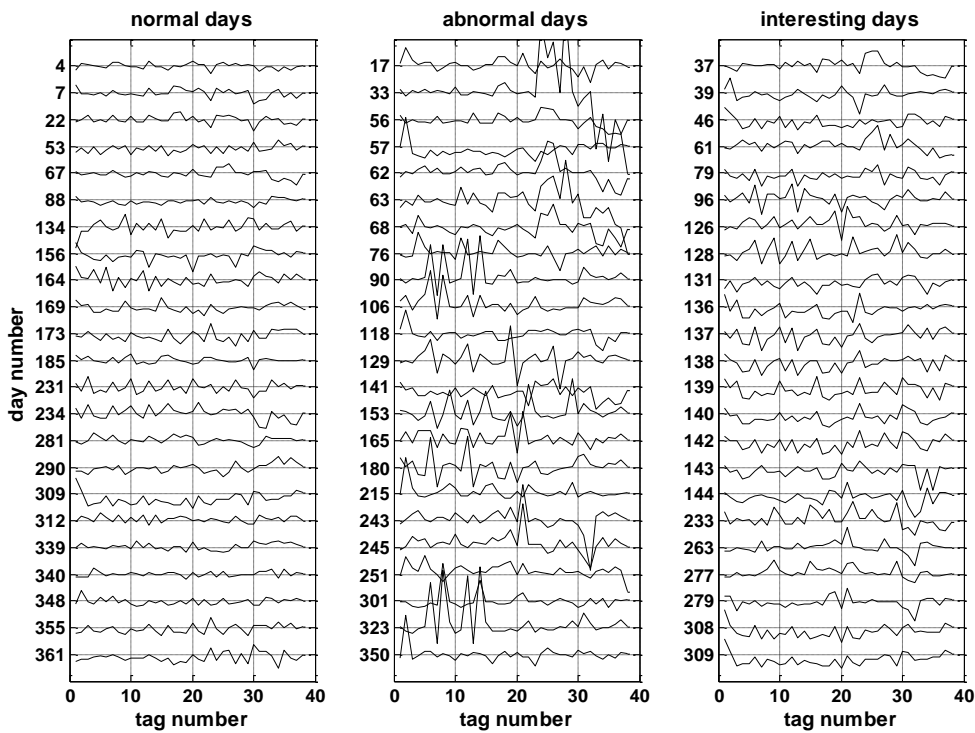


Figure 2. The plant profiles for selected days from the benchmark wastewater data set.

## 4.2 Results from process performance analysis

*PCA analysis:* PCA for process performance analysis required 10 components from the average eigenvalue criterion. The ten loading vectors are shown in Figure 3 and two-dimensional score plots are in Figure 4.

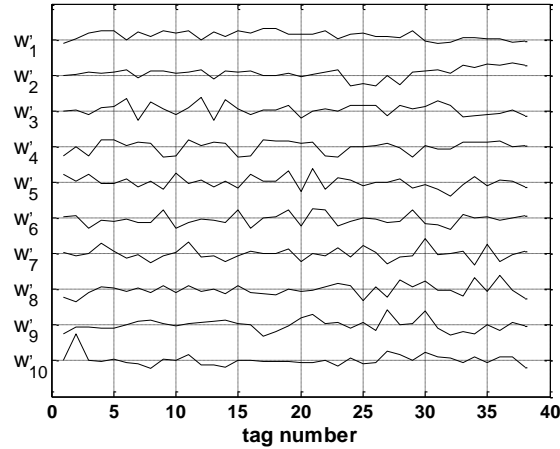


Figure 3. Loading vectors for process performance analysis

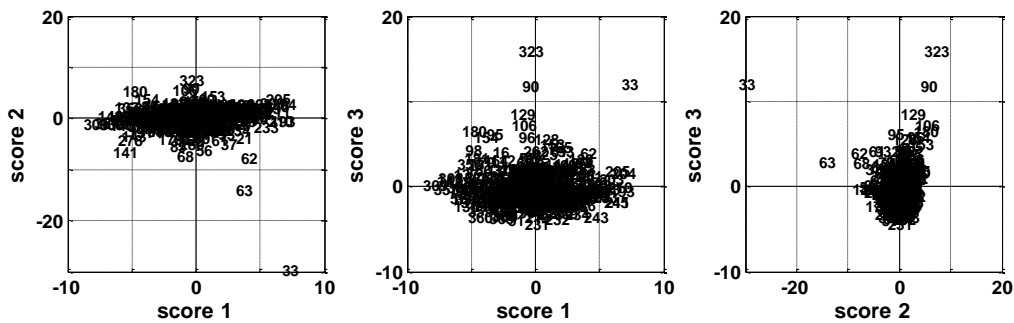


Figure 4. Two-dimensional score plots for process performance analysis

*Cluster and structure detection:* Figure 5 shows the hierarchical classification tree for the 10-dimensional score space of the plant performance analysis such that the right hand side of the upper panel joins onto the left side of the lower panel. The items in the tree represent the plant profiles on each day of operation and the vertical axis shows the angular separations. Days that are joined by horizontal lines low on the vertical axis have small angular separation in the score space and have similar plant profiles. The tree has been pruned to a manageable size by using the days with the 20% largest  $t'$  – vectors. Significant clusters reported by the automated clustering algorithm are indicated by black crosses on the horizontal axis.

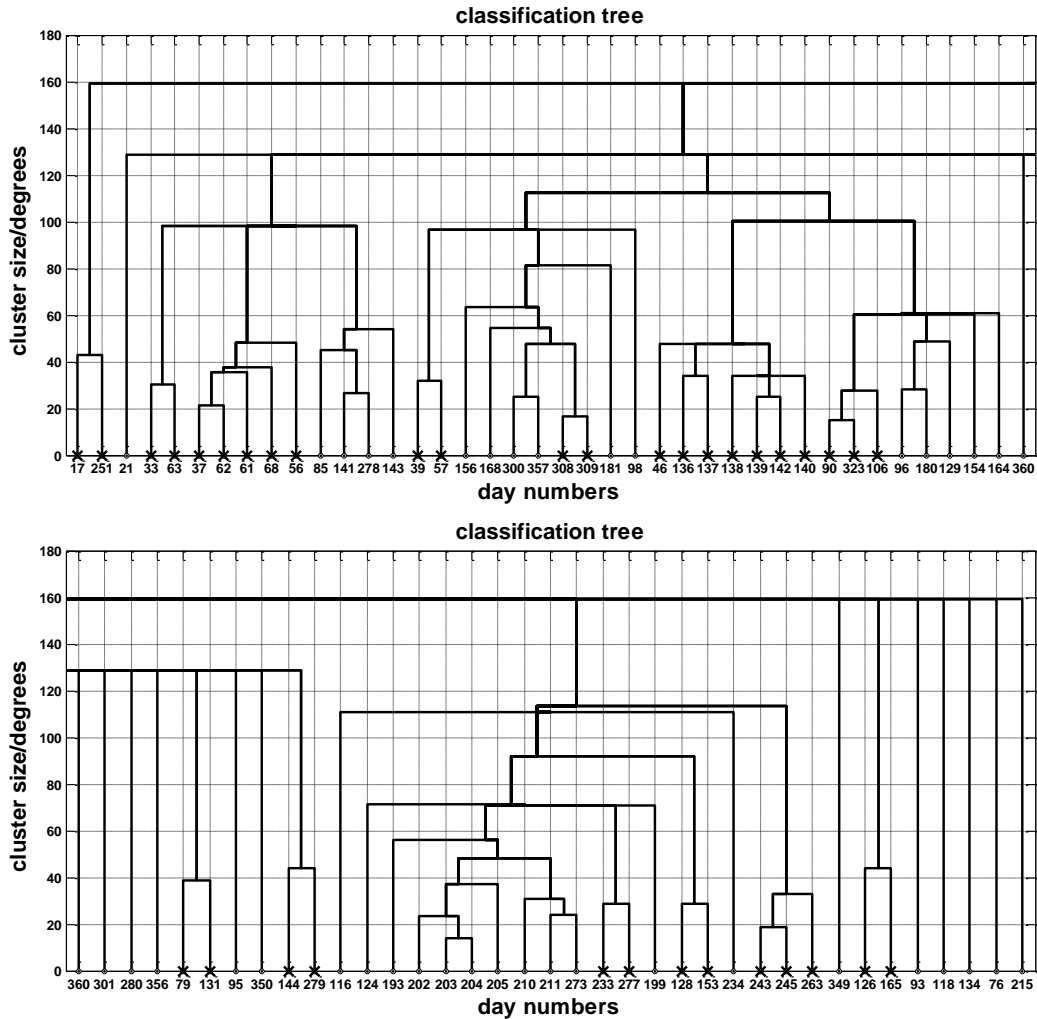


Figure 5. Hierarchical classification tree of the 10-dimensional score space in process performance analysis. The right side of the upper panel joins on to the left side of the lower panel. Crosses indicate clusters found by the automated algorithm.

The information accompanying the data set indicates unusual operating conditions on certain days which are listed in the fifth column of Table 1. Some detailed observations from the hierarchical tree are:

- Days 33 and 63 with secondary settler problems had similar plant profiles and were clustered together, as expected;
- Days 323 and 90 with solids overload problems are in a cluster. Day 106 also appeared in the same cluster and therefore it is likely that there was a solids overload problem on day 106 also;
- Days 153 and 180 with storms are not clustered together and therefore do not have similar plant profiles. However, the tree shows day 128 had a similar plant profile to day 153 suggesting it may also have been affected by a storm like the one on day 153;
- Days 243 and 245 and 263 have very similar plant profiles and therefore the operating conditions on those days were related;

- Some days are unique, for example day 350 is separated by a large distance from all other days in the data set. The reason for this is that day 350 had a unique profile with positive deviations in the value of Tag 2 and negative deviations in Tags 16 and 23, as can be seen in Figure 2;
- There are several other clusters, e.g. 46, 136-140 and 142. The right hand panel in Figure 2 shows the profiles of these days where it can be seen that 46, 136-140 and 142 have similar profiles that are different from the profiles on any other days. It would be necessary to go back to the operators' logs to find out what was special about the operation during those times.

These results show that the proposed methods can detect and display the structure within a high-dimensional score space.

*Detection of abnormal days:* Table 1 shows the automated outlier indexes. The analysis found the abnormal days previously reported in [40] and in [7] where manual inspection of a parallel coordinate plot was used. The outlier index also indicates additional days as abnormal that were not detected in previous studies. Day 350 is an example, and inspection of Figure 2 suggests that its classification as abnormal is correct because the plant profile for day 350 has an unusual and unique pattern of deviations.

The left panel of Figure 6 shows the outlier indexes in descending order and shows the cut-off threshold for abnormal operation. The outliers are on the very steep part of the plot while days classed as normal are in the region with a gentle slope. The right hand panel shows the abnormal days in Table 1 on a parallel coordinate plot. Each piecewise linear trend in the parallel coordinate plot shows the scores for one day's plant profile. Score values are on the vertical axis and the principal component number is on the horizontal axis. The grey lines are for normal days and the black lines are the days from Table 1 whose outlier indexes are larger than 1.

All the abnormal days identified in Table 1 also appear in the tree. Some, e.g. days 90, 323 and 106, are clustered together. Some, such as day 153, are in a cluster with another day or days that were not abnormal and some of the abnormal days are unique and not in any cluster (e.g. day 350). These results show that the hierarchical tree gives additional information over and above the algorithm for detection of abnormal days.

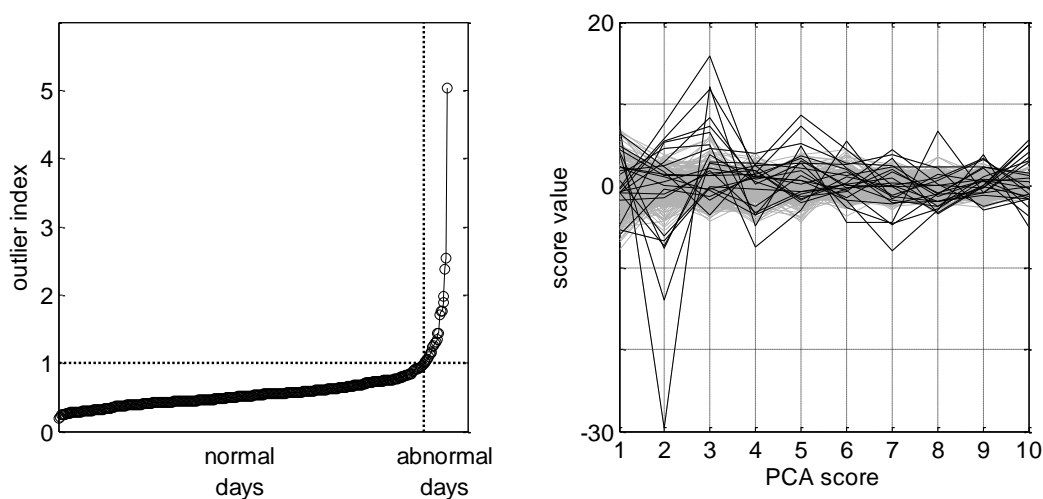


Figure 6. Automated outlier detection.

day number	date	outlier index	comments	known problem [40]	manual classification [7]
33	13/03/1990	4.95	in cluster {33 63}	secondary settler	clearly abnormal
323	28/05/1991	2.44	in cluster {90 106 323}	solids overload	clearly abnormal
63	29/04/1990	2.34	in cluster {33 63}	secondary settler	clearly abnormal
350	09/07/1991	1.98	unique		<i>not detected</i>
90	05/06/1990	1.82	in cluster {90 106 323}	solids overload	clearly abnormal
215	19/12/1990	1.76	unique		likely abnormal
57	22/04/1990	1.76	in cluster {57 39}		<i>not detected</i>
243	29/01/1991	1.70	in cluster {243 245 263}		clearly abnormal
245	31/01/1991	1.43	in cluster {243 245 263}		likely abnormal
251	07/02/1991	1.43	in cluster {17 251}		borderline
153	14/09/1990	1.30	in cluster {153 128}	storm	clearly abnormal
129	25/07/1990	1.29			likely abnormal
62	27/04/1990	1.28	in cluster {37 56 61 62 68}		likely abnormal
301	29/04/1991	1.23	unique		clearly abnormal
68	06/05/1990	1.23	in cluster {37 56 61 62 68}		<i>not detected</i>
141	10/08/1990	1.12			<i>not detected</i>
106	26/06/1990	1.12	in cluster {90 106 323}		likely abnormal
76	18/05/1990	1.12	unique		borderline
118	10/07/1990	1.08	unique		borderline
56	19/04/1990	1.02	in cluster {37 56 61 62 68}		<i>not detected</i>
165	03/10/1990	1.02	in cluster {126 165}		likely abnormal
180	22/10/1990	1.01		storm	clearly abnormal
17	29/01/1990	1.01	in cluster {17 251}		<i>not detected</i>
60	22/04/1990	< 1.0	not in the tree		likely abnormal

Table 1. Details of days classified as abnormal with comments on their analysis

### 4.3 Results from process audit during normal operations

*PCA analysis:* The PCA analysis for plant audit was conducted on the data set shown in Figure 7. All abnormal days have been removed because the aim is to give insights into the operation of the process during normal running. The reason why the time histories of some measurements look different than in Figure 1 (e.g. Tag 38) is that the time histories were rescaled after the abnormal days were removed, so more details can be seen. The data matrix has 38 rows, one for each measurement, and 357 columns, one for each day of normal operation in the data set. The PCA analysis required 11 components according to the average eigenvalue test. More principal components are needed after the abnormal days have been removed because the variability in the data set is more evenly spread when only the normal days are included<sup>2</sup>.

<sup>2</sup> For instance, in the hypothetical case that the time histories of an infinite number of normal days were identically distributed random variables then the average eigenvalue criterion would require all 38 components because all the eigenvalues of the variance matrix would all be equal.



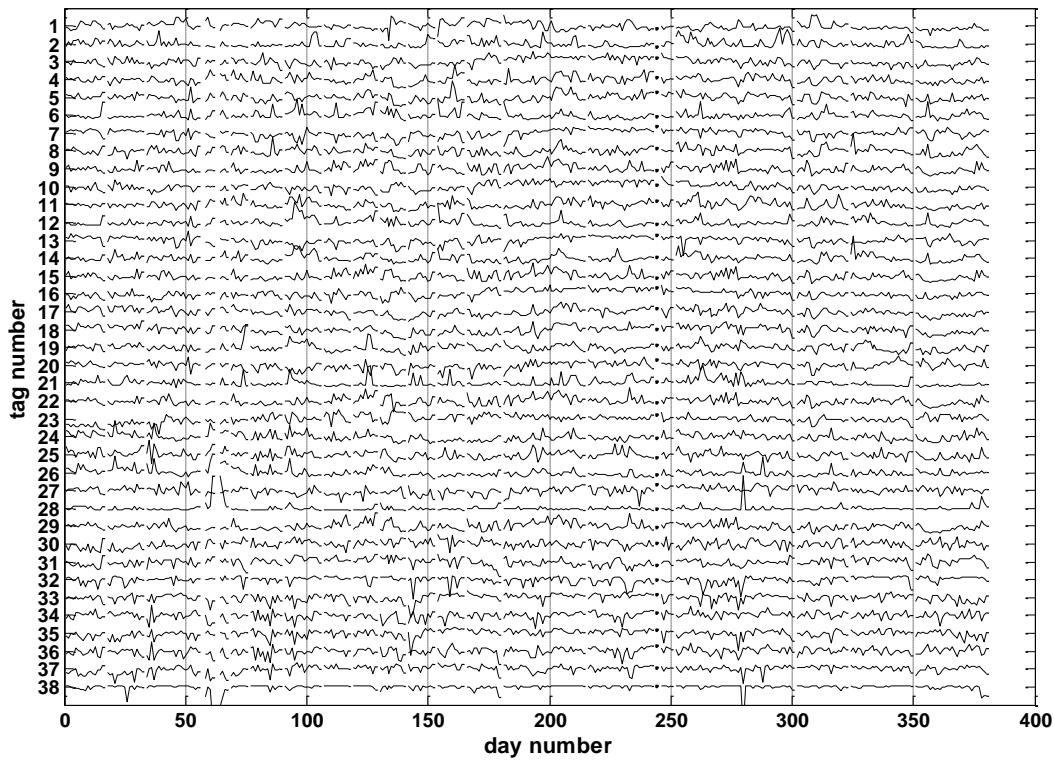


Figure 7. Daily averages of 38 measurements from the benchmark wastewater data set excluding abnormal days.

*Cluster and structure detection:* Figure 8 shows the hierarchical classification tree for the 11-dimensional score space in which each item represents one tag. The time history in the top trend in Figure 7 is represented by the spot labelled 1 on the left side of Figure 8. It was not necessary to prune the tree because the number of items is small enough for good visualization. Significant clusters detected by the automated clustering algorithm are highlighted with black crosses on the horizontal axis. Table 2 lists the tags lying in significant clusters and gives their descriptions. Process insights arising from the clusters include:

- Cluster 1 comprising Tags 9, 15, 22, and 29 shows that the wastewater process has little impact on conductivity because the input, output and intermediate conductivities form a very tight cluster. The conclusion is that the daily averages of these measurements move in a co-ordinated fashion. Any one measurement could act as a proxy for the others if a conductivity sensor were to fail;
- Cluster 6 (Tags 3, 10, 16, 23) shows the variations in pH measurements throughout the process are similar to one another and not strongly correlated with any other measurement. However, the output pH (tag 23 on the far right of Figure 8) is separated by about 45 degrees from the rest of the pH measurements suggesting that, while the influence of input pH does propagate through the plant, it becomes less pronounced as it propagates;
- Similar comments apply to the volatile suspended solids cluster 9 (Tags 7, 13, 20, 27);

- The separations of clusters 5 (6 and 12) and 4 (19 and 21) show the primary settler has a significant impact on suspended solids because the daily averages of the suspended solids to the secondary settler (19 and 21) is not in a cluster with the suspended solids to the primary settler (6 and 12);
- Cluster 11 (24, 26, 28) shows that variations in the output biological demand for oxygen clusters match those of the output solids and output sediment. Tag 25 is not a member of this cluster. This may be a significant observation because Tag 25 is the output chemical demand of oxygen. The output biological and chemical demands for oxygen are therefore shown by the process audit to be not closely related.

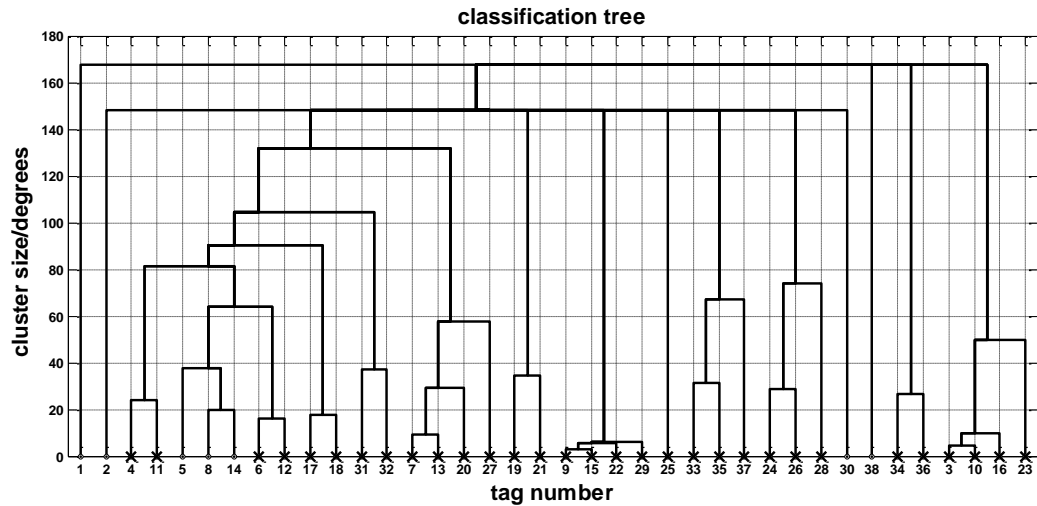


Figure 8. Hierarchical classification tree for the 11-dimensional score space in process audit. Crosses indicate the clusters of similar tags reported by the automated algorithm.

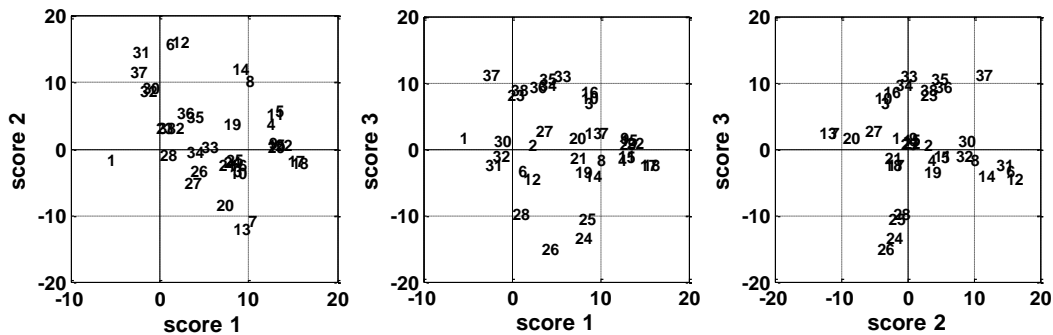


Figure 9. Two-dimensional score plots for process audit

<b>Cluster 1</b>	<b>Cluster 7</b>
9 input conductivity to plant	4 input Biological demand of oxygen to plant
15 input conductivity to primary settler	11 input Biological demand of oxygen to primary settler
22 input conductivity to secondary settler	<b>Cluster 8</b>
29 output conductivity	31 performance input suspended solids to primary settler
<b>Cluster 2</b>	32 performance input sediments to primary settler
34 performance input chemical demand of oxygen to secondary settler	<b>Cluster 9</b>
36 global performance input chemical demand of oxygen	7 input volatile suspended solids to plant
<b>Cluster 3</b>	13 input volatile suspended solids to primary settler
17 input Biological demand of oxygen to secondary settler	20 input volatile suspended solids to secondary settler
18 input chemical demand of oxygen to secondary settler	27 output volatile suspended solids
<b>Cluster 4</b>	<b>Cluster 10</b>
19 input suspended solids to secondary settler	33 performance input Biological demand of oxygen to secondary settler
21 input sediments to secondary settler	35 global performance input Biological demand of oxygen
<b>Cluster 5</b>	37 global performance input suspended solids
6 input suspended solids to plant	<b>Cluster 11</b>
12 input suspended solids to primary settler	24 output Biological demand of oxygen
<b>Cluster 6</b>	26 output suspended solids
3 input pH to plant	28 output sediments
10 input pH to primary settler	
16 input pH to secondary settler	
23 output pH	

Table 2. Clusters of tags with similar time histories identified in process audit

*Statistical interpretation of the hierarchical tree:* As was indicated earlier, in a process audit application  $\cos(\theta_{i,j})$  is the correlation coefficient between the  $i$ 'th and  $j$ 'th rows of  $\mathbf{X}_p$  (equation (8)). This section examines the correlations to give an insight into the clusters in the hierarchical tree.

The upper panel of Figure 10 shows a hierarchical tree calculated directly from  $\arccos(\mathbf{R})$  where  $\mathbf{R}$  is the correlation coefficient matrix for the rows of the data matrix  $\mathbf{X}$ . Equivalently, the same tree can be created from principal component analysis in which all 38 principal components are used. The tree in the upper panel has fewer significant clusters than the optimized tree in Figure 8, and some clusters have fewer tags. For instance Tag 27 is missing from the suspended solids cluster (7 13 20 27) and the separation of Tags 7 and 13 is more than 30 degrees instead of 10 degrees.

The white circles in the left panel of Figure 11 give a scatter plot of the daily averages of Tag 7, (the 7<sup>th</sup> time history in Figure 7) versus the daily averages of Tag 13. Each circle represents one day of operation. The elongated shape shows correlation exists but there is also noise and scatter in the measurements. The black diamonds are a scatter plot for the 7<sup>th</sup> versus 13<sup>th</sup> rows of  $\mathbf{X}_p$  in which 11 principal components have been used in the reconstruction. Much of the noise has been removed because only significant principal components were retained and the correlation therefore shows up more prominently resulting in a tighter cluster in the hierarchical tree.

The lower panel of Figure 10 shows the other extreme of a hierarchical tree where only three principal components are in use. The reason for demonstrating the case with three PCs is to show that the three-panel score plot (Figure 9) which visualizes the first three PCs only can be misleading.

The tree in the lower panel appears to have a cluster involving tags 24, 25, 26 and 28. These tags can also be seen lying close together in the right hand score plot of Figure 9. The inclusion of Tag 25 in this cluster is spurious, however. Tags 24, 25 are widely separated in the optimal 11-dimensional PCA score space but have projected onto adjacent spots in the Score 2-Score 3 plane. The white circles in the scatter plot in the right hand panel of Figure 11 represent the daily averages in the 24<sup>th</sup> and 25<sup>th</sup> rows of  $\mathbf{X}$  while the small grey squares are the 24<sup>th</sup> and 25<sup>th</sup> rows of  $\mathbf{X}_p$  using eleven PCs. The grey squares are generally close to the original data showing that the both measurements are well modelled by eleven principal components. By contrast, the black diamonds of the three PC model lie on a straight line and are not close to the original data. The three PC model gives a spurious correlation by ignoring significant information in principal components 4 to 11. This example shows the benefit of using multi-dimensional visualization of the optimized score space to avoid false conclusions.

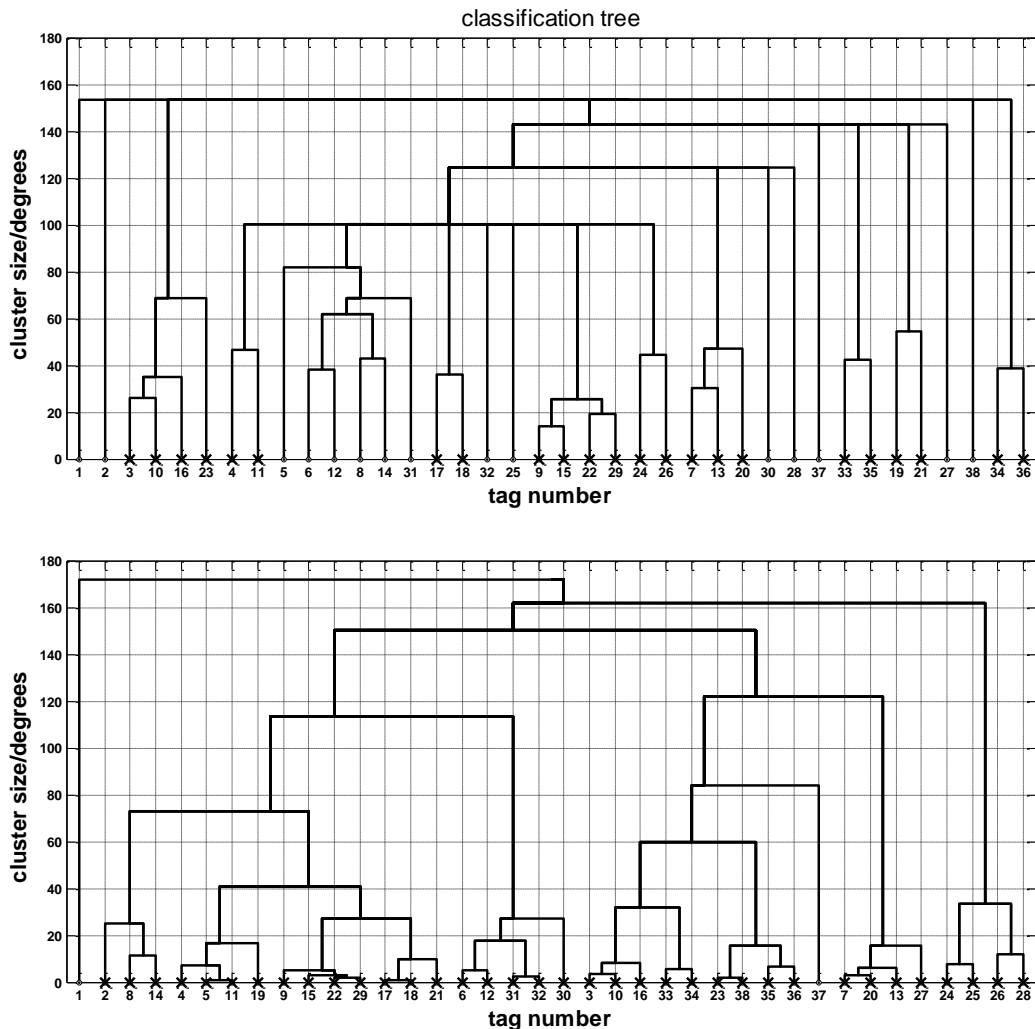


Figure 10. Hierarchical classification trees for process audit. Upper panel: using the full data matrix (all principal components). Lower panel: using three principal components.

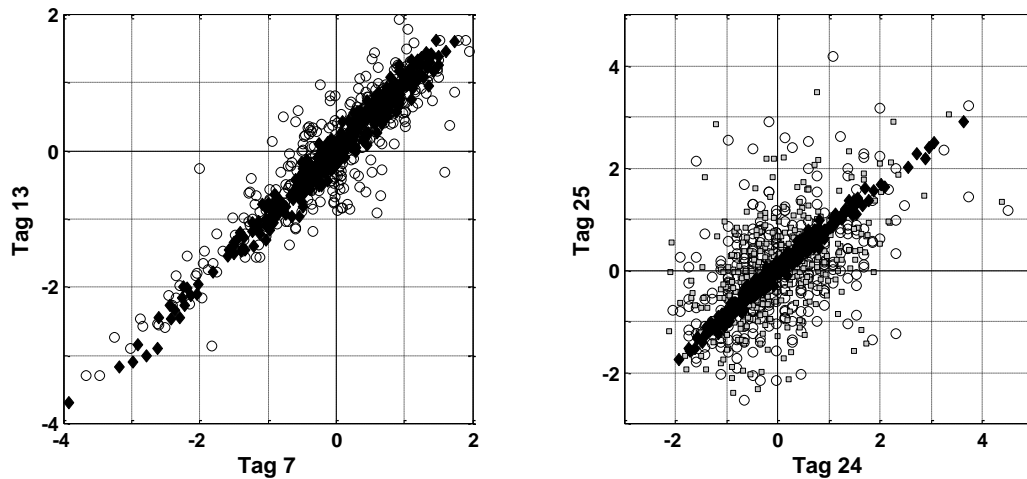


Figure 11. Scatter plots of daily averages. Left panel: Tag 7 versus Tag 13 using original data (circles) and 11 principal components (black diamonds). Right panel: Tag 24 versus Tag 25 using original data (circles), 11 PCs (grey squares) and three PCs (diamonds).

## 5 Case study 2 – compressor performance

*The data set*: Figure 12 shows a data set of hourly averages from the operation of a compressor. The first five tags are measurements of the flow, inlet and outlet pressure and temperature. The remaining tags labelled as SS1 to SS7 are soft-sensor estimates of quantities related to condition and efficiency. These are derived from the measurements, the compressor curves and thermodynamical principles. Figure 13 shows a selection of the plant profiles at various hours.

The case study involves operation at different set points and has been included to demonstrate the use of Euclidian distances in a process performance analysis for the classification of different operating states.

*PCA analysis*: The PCA analysis required 4 components from the average eigenvalue criterion. The two-dimensional score plots are in Figure 14, however they do not show the full model because four principal components were needed. The nature of the plots guides the selection of parameters for the hierarchical tree, however.

There are three main clusters separated by their score 1 values, as can be seen in the left hand score plots. One is at the origin, the others have positive and negative values of score 1. This is to be expected because in Figure 12 the operating states are above average at the start of the data set, average in the middle and below average towards the end. The clusters in the score plot do not form plumes radiating from the origin and therefore a Euclidian distance measure is appropriate for the hierarchical tree.

There are many hours of operation in each operating state and the tree can be pruned to enhance visibility using every 40<sup>th</sup> plant profile rather than all of them. The plant performance analysis therefore uses the 50 plant profiles from Figure 13.

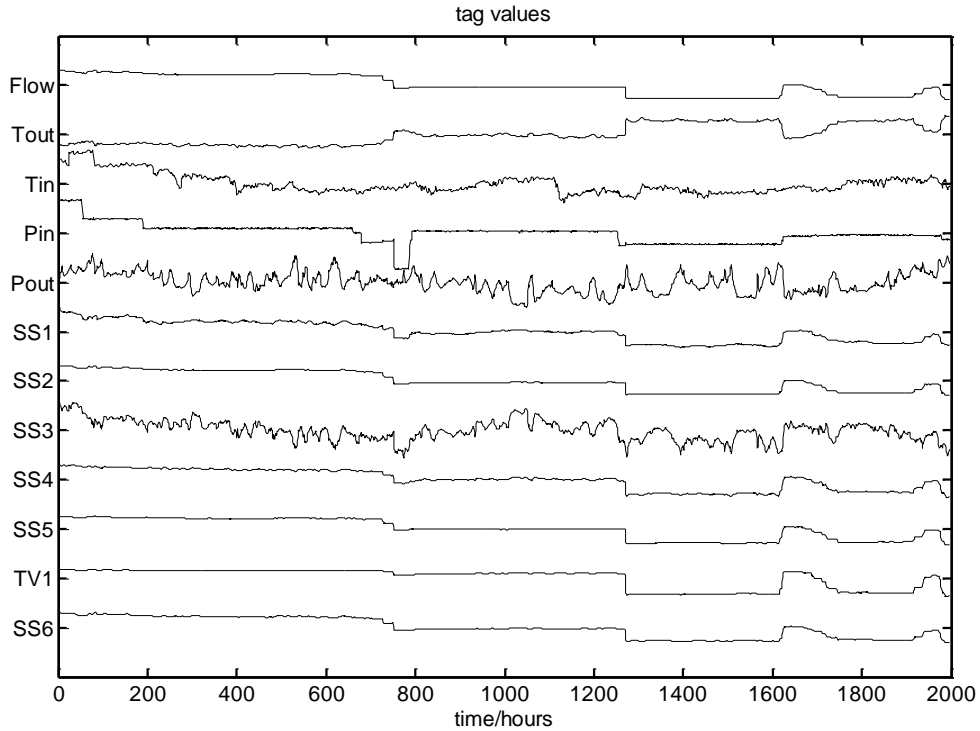


Figure 12. Hourly averages from operation of a compressor

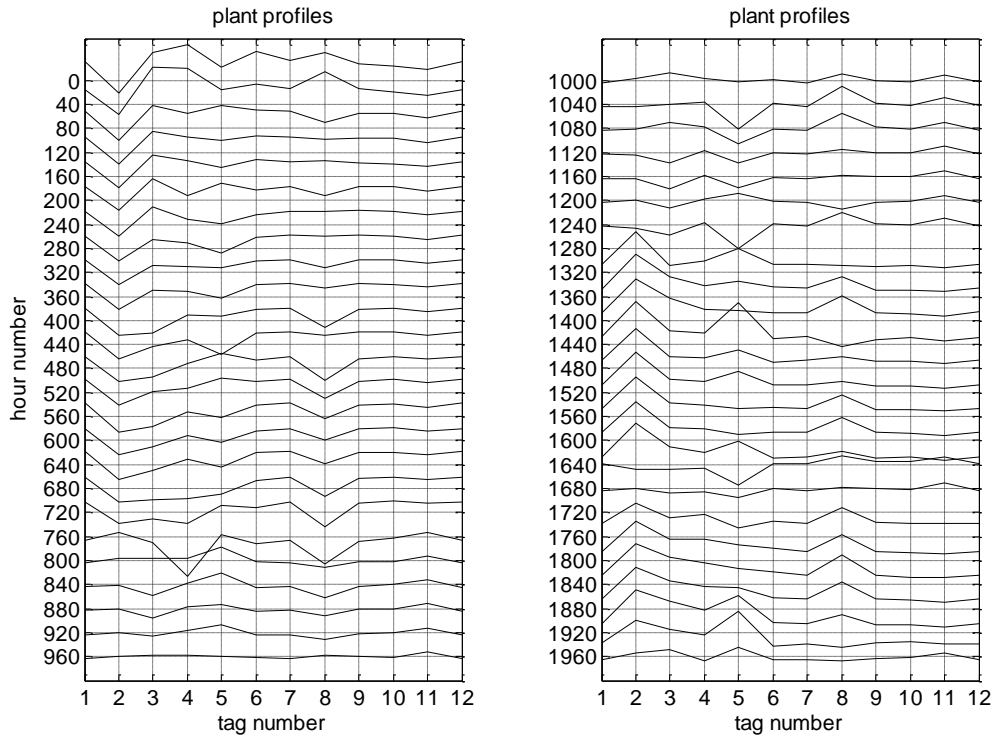


Figure 13. A subset of the plant profiles taken every forty hours.

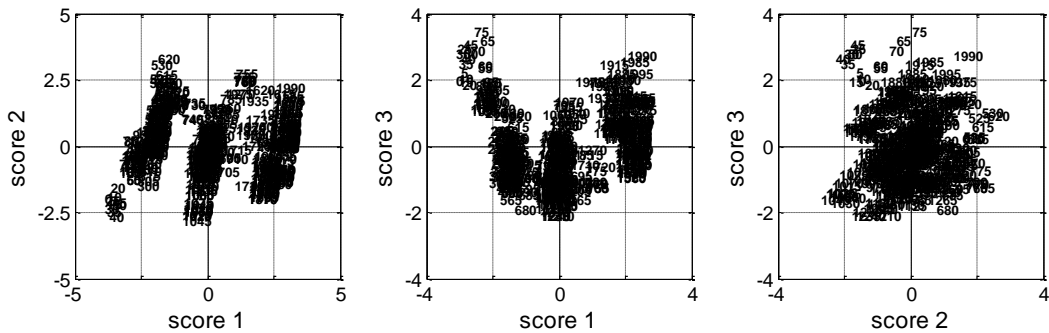


Figure 14. Two-dimensional score plots for process performance analysis

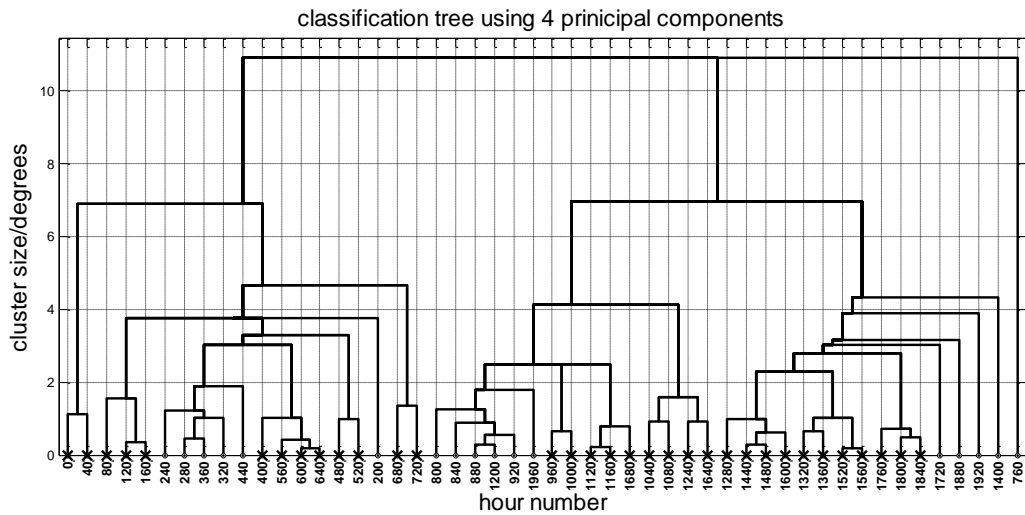


Figure 15. Hierarchical classification tree for process performance analysis

*Cluster and structure detection:* Figure 15 shows the hierarchical tree. Features of the performance are highlighted by the tree. On the right hand side, it shows that hour 760 had a plant profile that was greatly different from any other profile in the data set. Figure 12 shows the  $P_{in}$  measurement was abnormally low at that time. The tree also shows that hours 1 and 40 at the start of the data set are not like any other hours of operation.

By visual observation, the tree shows three main groups representing the three main operating states. Broadly, the left hand cluster covers hours 80 to 720 (operation above average values), the middle cluster covers hours 800 to 1240 together with 1640, 1680 and 1960 (operation at average values), while the right hand cluster covers hours 1280 to 1920 where the operating values were below average. The same groups can be found by choosing a branch length ratio of 0.67 in the significant clusters algorithm.

An analysis using a branch length ratio of 1 reveals the significant clusters represented by crosses in Figure 15. These show the detailed structure within the three main areas of operation. One finding is

that the plant profile at hour 1680 was similar to the profiles at hours 1120 and 1160 while hour 1640 is in a sub-cluster with 1040, 1080 and 1240 (these sub-clusters are near the centre of the tree and slightly to the right). The significance of this observation is that it shows the operation was in a transitional state at hours 1680 and 1640 since these are more similar to some much earlier hours of operation than to each other. A further observation is that hours 80, 120 and 160 (on the left side of the tree) are somewhat separated. The reason for their separation is that  $P_{in}$  had a small step down to a new constant value at about hour 180.

## 6 Conclusions

A new method of visualization of the structure within a high-dimensional principal component analysis was presented using an agglomerative hierarchical classification tree. Its applications include performance monitoring and the auditing of the condition of chemical processes and equipment. Performance monitoring is a PCA analysis in which the items of interest are the plant profiles at different times, for instance hourly or daily averages. Process audit means that the items in the analysis are the time histories of the plant measurements.

A classification tree is based on a measure of similarity between items. Two measures were considered, one was the angular separation of the items in the multidimensional score space, the other was the Euclidian distance between them. The reasons for using one measure or the other were discussed and both were demonstrated.

Two automated algorithms were presented. One determined the structure of the tree and gave a text-format report from which the x-y coordinates for the plotting of the tree could be determined. The report could also be parsed to determine the items belonging to significant clusters. The other algorithm detected outliers using a concept introduced by Wang *et. al.*, [7], who treated an item having an extreme value in any one score as abnormal. When applied to a benchmark data set the automated outlier detection gave the same results as those found manually in [7] and they matched the original observations about operating conditions in [40]. Additional information was generated from the clusters detected in the hierarchical tree, in particular additional days of operation were detected that had plant profiles similar to but less extreme than the abnormal days, suggesting that the same operational difficulties were present on those days but to a lesser extent.

Finally, some instances were highlighted where two-dimensional score plots gave misleading results. The ambiguities were resolved in the hierarchical tree which clearly showed the benefit of displaying the high-dimensional structure in a tree plot.

## 7 Acknowledgements

The first author gratefully acknowledges the support of the Royal Academy of Engineering (Global Research Award) and of ABB Corporate Research. The authors thank Statoil Tjeldbergodden for permission to use the compressor data set.

## 8 References

1. Wise, B.M., and Ricker, N.L., Veltkamp, D.F., and Kowalski, B.R. (1990). A theoretical basis for the use of principal components models for monitoring multivariate processes, *Process Control and Quality*, 1, 41-51
2. Kresta, J.V., MacGregor, J.F., and Marlin, T.E. (1991). Multivariate statistical monitoring of process operating performance, *Canadian Journal of Chemical Engineering*, 69, 35-47.
3. Wise, B.M., and Gallagher, N.B. (1996). The process chemometrics approach to process monitoring and fault detection, *Journal of Process Control*, 6, 329-348



4. Kourti T, 2005, Application of latent variable methods to process control and multivariate statistical process control in industry, *International Journal of Adaptive Control and Signal Processing*, 19, 213-246.
5. Qin S.J. 2003, Statistical process monitoring: basics and beyond, *Journal of Chemometrics* 17, 480-502.
6. Antaki, J., Paden, B.E., Piovoso, M.J., and Banda, S.S., 2002, Award-winning control applications, *IEEE Control Systems Magazine (Dec)*, 22, 8-20,
7. Wang, X.Z., Medasani, S., Marhoon, F., and Albazzaz, H., 2004, Multidimensional visualization of principal component scores for process historical data analysis, *Industrial Engineering and Chemistry Research*, 43, 7036-7048.
8. Chatfield, C., and Collins, A.J., 1980, *Introduction to multivariate analysis*, Chapman and Hall, London, UK.
9. Wold, S., Esbensen, K., and Geladi, P., 1987, Principal Component Analysis, *Chemometrics and Intelligent Laboratory Systems*, 2, 37-52.
10. Alam. T.M., and Alam, M.K., 2005, Chemometric analysis of NMR spectroscopy data: A review, *Annual Reports on NMR Spectroscopy*, 54, 41-80.
11. Ozaki, Y., Sasic, S., and Jiang, H.H., 2001, How can we unravel complicated near infrared spectra? Recent progress in spectral analysis methods for resolution enhancement and band assignments in the near infrared region, *Journal of Near Infrared Spectroscopy*, 9, 63-95.
12. Seasholtz, M.B., 1999, Making money with chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 45, 55-64.
13. Tzeng, D.Y., and Berns, R.S., 2005, A review of principal component analysis and its applications to color technology, *Color Research and Application*, 30, 84-98.
14. De Belie, N., De Smeldt, V., and De Baerdemaeker, J., 2000, Principal component analysis of chewing sounds to detect differences in apple crispness, *Postharvest Biol. Technol.*, 18., 109-119.
15. Brodnjak-Voncina, D., Dobcnik, D., Novic, M., and Zupan, J., 2002, Chemometrics characterisation of the quality of river water, *Analytica Chimica Acta*, 462,87-100.
16. Bengraïne, K., and Marhaba, T.F., 2003, Using principal component analysis to monitor spatial and temporal changes in water quality, *Journal of Hazardous Materials*. 100, 179-195.
17. Haag, I., and Westrich, B., 2002, Processes governing river water quality identified by principal component analysis, *Hydrological Processes*, 16, 3113-3130
18. Wouters, L., Gohlmann, H.W., Bijmens, L., Kass, S.U., Molenberghs, G., and Lewi, P.J., 2003, Graphical exploration of gene expression data: A comparative study of three multivariate methods, *Biometrics*, 59, 1131-1139.
19. Wu, H.D., Siegel, M., and Khosla, P., 1999, Vehicle sound signature recognition by frequency vector principal component analysis, *IEEE Transactions on Instrumentation and Measurement*, 48 1005-1009
20. Malhi. A., and Gao, R.X., 2004, PCA-based feature selection scheme for machine defect classification, *IEEE Transactions on Instrumentation and Measurement*, 53, 1517-1525.
21. Flaten, G.R., Belchamber, R., and Collins, M., and Walmsley, A.D., 2005, Caterpillar - an adaptive algorithm for detecting process changes from acoustic emission signals, *Analytica Chimica Acta*, 544, 280-291.
22. Belchamber, R.M., and Collins, M.P., 1993, *Method for monitoring acoustic emissions*, European Patent Office, Publication N<sup>o</sup> 0 317 322 B1.
23. Goulding, P.R., Lennox, B., Sandoz, D.J., Smith, K.J., and Marjanovic., O., 2000, Fault detection in continuous processes using multivariate statistical methods, *International Journal of Systems Science*, 31, 1459-1471.
24. Kourti, T., and MacGregor, J.F., 1996, Control of multivariate processes, *Journal of Quality Control*, 28, 409-428
25. Jackson, J.E., and Mudholkar, G.S., 1979, Control procedures for residuals associated with principal components analysis, *Technometrics*, 21, 341-349.
26. Martin, E.B., and Morris, A.J., 1996, Non-parametric confidence bounds for process performance monitoring charts, *Journal of Process Control*, 6, 349-358.
27. Duda, R.O., Hart, P.E., Stork, D.G., 2000, *Pattern Classification (2nd Edition)*, Wiley-Interscience.
28. Oja E, 2004, Finding clusters and components by unsupervised learning, *Lecture Notes in Computer Science*, 3138: 1-15.
29. Gordon, A.D., 1987, A review of hierarchical classification, *Journal of the Royal Statistical Society*, 150, 119-137.
30. Seem, J.E., 2005, Pattern recognition algorithm for determining days of the week with similar energy consumption profiles, *Energy and Buildings*, 37, 127-139.
31. Norris, T., Paul K. Aldridge, P.K., and Sekulic, S.S., 1997, Determination of end-points for polymorph conversions of crystalline organic compounds using on-line near-infrared spectroscopy, *Analyst*, 122, 549-552.
32. Wiedemann, L.S.M., d'Avila, L.A., and Azevedo, D.A., 2005, Adulteration detection of Brazilian gasoline samples by statistical analysis, *Fuel*, 84, 467-473.
33. Hwang, D-H., Han, C., 1999, Real-time monitoring for a process with multiple operating modes, *Control Engineering Practice*, 7 891-902.
34. Lee, Y-H., Min, K.G., Han, C., K Chang, K.S., and Choi, T.H., 2004, Process improvement methodology based on multivariate statistical analysis methods, *Control Engineering Practice*, 12, 945-961.
35. Valle, S., Li, W.H., and Qin, S.J., 1999, Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods, *Industrial Engineering and Chemistry Research*, 38, 4389-4401.
36. Raich, A., and Çinar, A., 1997, Diagnosis of process disturbances by statistical distance and angle measures, *Computers & Chemical Engineering*, 21, 661-673.
37. Johannesmeyer, M.C., Singhal, A., and Seborg, D.E., 2002, Pattern Matching in Historical Data, *AIChE Journal*, 48, 2022-2038.

38. Yang, J., Ward, M.O., and Rundensteiner, E.A., 2003, Interactive hierarchical displays: a general framework for visualization and exploration of large multivariate data sets, *Computers & Graphics*, 27, 265–283
39. Faults in a urban waste water treatment plant, maintained by University of California, Irvine, CA. [Online] <http://www.ailab.si/orange/doc/datasets/water-treatment.htm>
40. Sánchez, M., Cortés, U., Béjar, J., Gracia, J., D.; Lafuente J., and Poch, M., 1997, Concept formation in WWTP by means of classification techniques: a compared study, *Applied Intelligence*, 1997, 7, 147-165.

## Appendix: Worked example

Reporting and classification – worked example: An example of the report generated by the automated algorithm in the process performance analysis of Figure 5 are as follows. These steps show the building up of the cluster comprising days 138, 139, 140 and 142:

<u>Iteration 11:</u>	<u>Iteration 14:</u>	<u>Iteration 16:</u>
clustersize: 24.9	clustersize: 34.2	clustersize: 34.2
current: (139 142)	current: (138 139 142)	current: (138 139 142 140)
left: 139	left: 138	left: (138 139 142)
right: 142	right: (139 142)	right: 140

The *current* heading shows the items in the current cluster while the *left* and *right* headings show the sub-clusters that have been joined to make the current cluster. The cluster size is the maximum out of all the  $\theta_{i,j}$  values for the items of the current cluster and in this case the size of the cluster did not grow at iteration 3 when day 140 was added. Table 3 shows why. Days 139 and 142 would be clustered first because the smallest angle is  $\theta_{139,142} = 24.9$ . The next day to join the cluster is 138 having an angular separation of 26.1 from day 142. The overall cluster size for the (138 139 142) cluster is  $34.2^\circ$  because 138 and 139 are separated by that angle. Day 140 then joins the cluster because it is the next closest with  $\theta_{140,142} = 27.1$ . The overall cluster size remains the same, however, because the maximum distance in the cluster is still the  $34.2^\circ$  between days 138 and 139.

Significant cluster detection – worked example: Results from later iterations are:

<u>Iteration 37:</u>	clustersize: 47.7
	current: (46 136 137 138 139 142 140)
	left: 46
	right: (136 137 138 139 142 140)
<u>Iteration 46:</u>	clustersize: 100.2
	current: (46 136 137 138 139 142 140 90 323 106 96 180 129 154 164)
	left: (46 136 137 138 139 142 140)
	right: (90 323 106 96 180 129 154 16)

The clustersize shows that the horizontal line linking the (46 136 137 138 139 142 140) group is at 47.4 degrees on the vertical axis and that this group joins the (90 323 106 96 180 129 154 16) group to form a cluster of size 100.2. Therefore the branch for the (46 136 137 138 139 142 140) has a length of  $100.2 - 47.4 = 52.8$  degrees. The ratio between the branch length and the size of the (46 136 137 138 139 142 140) cluster is  $52.8/47.4 = 1.11$ , so the (46 136 137 138 139 142 140) cluster is classified as a significant cluster because the branch to cluster ratio is larger than one.

day	138	139	140	142
138	0	34.2	32.2	26.1
139	34.2	0	26.6	24.9
140	32.2	26.6	0	27.1
142	26.1	24.9	27.1	0

Table 3. Part of the angular distance matrix **A** used to create the tree in Figure 5