Pump leak Detection Using DPCA

Bahador Rashidi and Qing Zhao Department of Electrical and Computer Engineering Univ. of Alberta, Edmonton, Alberta, Canada

Date: Nov. 1, 2016

Abstract

This report is focused on fault detection in a pump system using dynamic prinsipal component analysis. DPCA is deployed to detect leakage malfunctions in 40 data-set. This report also present the detailed steps for determining tuning parameters for DPCA model. In this report, the result of leakage detection using ordinary PCA is compared with DPCA.

1. Dynamic Principal Component Analysis (PCA)

Dynamic principal component analysis (DPCA) is a tool that can be deployed for different purposes (e.g. fault detection, system identification, dimension reduction, *etc.*). There are some user defined tuning parameters in DPCA that have crucial roles for a particular application. This research used DPCA for fault detection in a pump system which is a mechanical process with relatively fast time-constant. In report phase one [1], the detail explanation about DPCA is provided and here only a brief overview about the DPCA and detail steps for tuning the user defined parameters are presented.

Consider $X_{Original} \in R^{N \times m}$ is the matrix of process variables such that columns are process variables and rows are sample observation values. In a majority of processes, the generic relation between process variable are dynamic and nonlinear. Therefore, ordinary PCA may not be appropriate monitoring tool. However, by some preparatory step on the original data-set $X_{original}$, PCA can be still used for fault detection purposes. The augmentation step that is indicated in fig. 1, makes PCA applicable for dynamic systems. Also, if there is a low level of non-linearity among process variables, DPCA fits the best linear approximation model to the process.

As shown in fig. 1, augmented matrix $X \in R^{(N-\tau \times h) \times (m \times (h+1))}$ is constructed from $X_{original}$. Two tuning parameters in this step are, first, τ that is the sample lag considered for shifting, and second, h that is the number shifting. The detail information about determining these two parameters are explained in section 2. After construction of augmented matrix, it is deemed as the main data-set matrix and PCA is conducted on it.

The underlying idea of the PCA is to extract the dominant principal components (PCs) of a data-set. SVD is a common tool to derive PCs that is also deemed in this report. For this phase of the project, PCA is considered as a mechanism to check the presence of abnormality in sensor and process variable measurement. To this end, First, matrix X needs to be mean-centered and scaled to unit-variance. Then by using SVD on $\Phi = \frac{1}{\sqrt{N-1}}X$ as (1), loading and score matrices

are determined.

$$\Phi = USV^T \tag{1}$$

S is a diagonal matrix including singular values of Φ arranged in descending order. V is orthonormal matrix that is called loading matrix of ϕ . The number of latent variables or principal components of matrix X is p_{com} that is the smallest number satisfies the inequality,

$$\frac{\sum_{i=1}^{p_{com}} \delta_i \times 100}{\sum_{j=1}^{j=(m+n)l} \delta_j} \ge CPV \quad p_{com} \le m \times (h+1)$$
(2)

where CPV is cumulative percentage value explained in section 2 and δ_i are singular values of Φ . Two corresponding loading vectors for matrix X can be defined as $\hat{P} = V_1 \in R^{p_{com} \times m}$ and $\tilde{P} = V_2 \in R^{(m-p_{com}) \times m}$. By using \hat{P} , the new data-set in principal and residual subspaces are $X_{PC} = X\hat{P}$ and $X_{RES} = X\tilde{P}$, respectively. Eq. (3) shows two projection matrices $\hat{\Pi}$ and $\hat{\Pi}$ that can decompose data matrix X into two parts aligned with PCs and residue subspace.

$$X = \hat{X} + \tilde{X} = X\hat{\Pi} + X\tilde{\Pi} = X\hat{P}\hat{P}^T + X\tilde{P}\tilde{P}^T$$
(3)

As can be concluded from (3), \hat{X} is the projection of original process data matrix X into the principal planes.

In literature, two T^2 Hotelling and square prediction error (SPE) are defined as indices for monitoring and changes in principal and residual subspaces, respectively. By calculating upper control limits (UCL) for both indices as their threshold for normal behavior of the process (no-fault situation), faults can be detected when mentioned thresholds (UCL) are violated. (4) and (5) show the formula for calculation of indices. The steps for calculation UCLs are explained in section 2.

$$T_{Hotelling}^{2} = ||\hat{X}||^{2} = X\hat{P}S^{-2}\hat{P}^{T}X$$
(4)

$$SPE = ||\tilde{X}||^2 = ((I - \hat{\Pi})X)^T ((I - \hat{\Pi})X)$$
 (5)

where, S is the variance of the scores.

In general, the first step for deploying DPCA is to train the base-line model. For that purpose, a suitable data-set which needs to have some feature is chosen. First, it most be clean and without any malfunctions. Second, it most include a relatively complete course of variation of the process variables. For instance in the pump data, the data-set number 5 is chosen since the input is not constant and change in course and reveals a relative complete mode of the system.

2. Tuning Parameters

Three user defined parameters of DPCA are τ which is the sample lag, h that is the number of shifting (augmentation, see fig. 1) and CPV which is the cumulative percentage value for determination of the principal components (latent variables).

2.1. CPV

Cumulative percentage value (CPV) is used to find the number of principal components or linear relations between the columns of the augmented matrix X. As the rule of thumb the CPV is chosen between 85 to 99.9 percent for different process. It should be noted that the number of linear relation between columns of the augmented matrix X is the total number of columns minus the p_{com} that is determined in Eq. 2.

2.2. *τ*

This parameters determines the number of sample lag required for each variable's shifting action. In order to determine the proper τ , information about the sampling frequency $(F_s = \frac{1}{T_s})$ of the measured variables and time constant $(T_c = \frac{1}{F_c})$ of the process is required. The F_s of the acquired data is always known. However, determining the time constant of the system might require a specific procedure in training data set (e.g. step test). The minimum value for the sampling frequency (F_s) with regards to Nyquist rule is $F_s = \frac{2}{T_c}$. Therefore, as a rule of thumb, the value of sample lag is chosen as.

$$\tau = \log\left(Floor\left[\frac{F_s}{F_c}\right]\right) \tag{6}$$

where, Floor is a function that find the floor round integer of its input.

2.3. h

This parameters determine the number of shifting action for process variables (see Fig. 1). Below is the algorithm for determining h.

- 1- Find the maximum sample delay (T_{delay}) between measured process variables and calculate the initial $h = Floor \left[\frac{T_{delay}}{\tau} \right]$.
- 2- set l = 0.
- 3- Construct augmented matrix X with corresponding sample delay τ and number of delay h = h + l.
- 4- Perform PCA and calculate the principal scores.
- 5- set $i = n \times h$ and r(l) = 0 where, n is number of process variables.
- 6- check if the i^{th} component represents a linear relation, if yes proceed, if no go to step 8.
- 7- set i = i 1 and r(l) = r(l) + 1, repeat step 6.

8- Calculate the number of new relationships between the column of the augmented matrix *X* as follows,

$$r_{new}(l) = r(l) - \sum_{j=0}^{l-1} (l-j+1) r_{new}(j)$$
(7)

- 10- Set l = l + 1 and go to step 3.
- 11- Set h = h + l and Stop.

For checking whether a component represent a linear relation, the corresponding singular value for that component most be among those that satisfied the inequality in Eq. 2.

2.4. Upper control limits

In order to determine the thresholds (upper control limit) for the indices in Eqs. (4 and 5), DPCA is conducted on the training data-set. The followings are the relations for these two indices.

$$UCL_{T^{2}} = \frac{p_{com}(N-1)(N+1)}{N^{2} - Np_{com}} F_{\alpha, p_{com}, N-p_{com}}$$
(8)

$$UCL_{SPE} = \theta_1 \left(\frac{c_{\alpha} \sqrt{2\theta_2 \beta^2}}{\theta_1} + \frac{\theta_2 \beta (\beta - 1)}{\theta_1^2} + 1 \right)^{1/\beta}$$
(9)

 $F_{\alpha,p_{com},N-p}$ is the upper α percentile of the F distribution with p_{com} and $N-p_{com}$ degree of freedom where, N is the number of observation for each variable. $\theta_i = \sum_{j=p_{com}+1}^{n(h+1)} \delta^i_j$ for i=1,2,3. After determining θ_i , β can be calculated as $\beta=1-\frac{2\theta_1\theta_3}{3\theta_2^2}$. c_α is standard normal variable corresponding to the upper $1-\alpha$ percentile. α is the confidence level that is expected for determined threshold that is sensitive to the sensor to noise ratio (SNR) of the measurements. the value of $\alpha \geq 95\%$ is commonly used otherwise the $SNR \leq 5$. More details about $F_{\alpha,p_{com},N-p}$ can be found in [2] and [3].

3. Identification of the Base-Line Model for Pump Data Using DPCA

In order to extract the base line model for the pump, one of the normal data sets needs to be considered. Amongst 33 provided data sets, data-set number 5 is chosen as the best candidate for representing the normal behavior of the pump. There are almost 20 captured data-sets for normal situation among provided data from industry, but data set number 5 is considered due to the fact that it covers more comprehensive mode of the pump in normal operating condition.

As explained in section 2, the CPV is chosen 98%. The average time constant for process variables is calculated 300 second and the sampling time for data acquisition is 5 second. Therefore, the sample lag for construction of the matrix X is determined $\tau = 1$. By following the steps in section 2.3, number of shifting original data matrix is calculated h = 15.

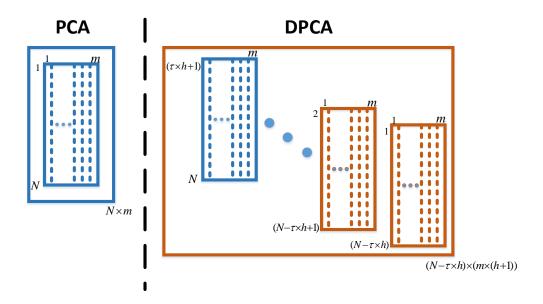


Figure 1. Schematic diagram of shifting data matrix x as a preparatory step of DPCA method

By doing this data shifting, each data-set $X = [x_i^T]$ becomes as $X_{shifted} = [x_i^T, ..., x_{i-15}^T]^T$ (see equation (10)). The concept of matrix augmentation that is the main difference between DPCA and PCA is shown in figure (1).

$$X_{shifted} = \begin{bmatrix} x_{15} & x_{14} & \dots & x_2 & x_1 \\ x_{16} & x_{15} & \dots & x_3 & x_2 \\ \dots & \dots & \dots & \dots \\ x_N & x_{N-1} & \dots & x_{N-14} & x_{N-15} \end{bmatrix}$$
(10)

For training data-set number 5, by considering cpv as 98%, 13 principal components are extracted out of total 60 components. It should be mentioned that the three parameters τ , h and cpv are empirical and need to be tuned for a process.

3.1. Results and discussion

all results for fault detection are presented in figures (3) to (34). As explained in the report phase one, T^2 and SPE indices are monitored for fault identification purposes. Each of these indices has an interpretation and with regards to the nature of the process one or both of them can be reflecting the presence of faults. T^2 statistic shows the variation of process variables' mean values in the main principal components and SPE index stands for the variation in the mean of the process variables in the residual subspace that is perpendicular to the main Pcs. In the given pump data, since we know that any variation of the process variables may not be because of the leakage and might be introduced to the system by input signal (VFD frequency), violation of the upper control limit (threshold) in the T^2 index plot does not infer to presence of fault(s) and means that the process variables are subjected to the changes in the main extracted PCs. Therefore, in order to detect leakage in the pump, users must observe the SPE index as the main plot representing the

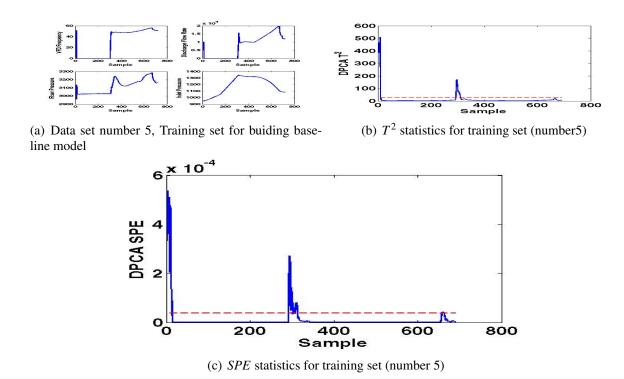


Figure 2. T^2 and SPE indices calculated by DPCA for normal training set (number5)

true status (no leak or leak) of the pump. For data sets number 24 to 34, both T^2 and SPE signals shows the presence of fault. However, in some case such as case number 10, 18 and 19 while threshold of T^2 statistic is exceeded, there is no leakage.

The summary of anomaly (leakage) detection using DPCA is brought in (3). Moreover, the obtained health monitoring results by the company's operators are compared with DPCA counter parts and brought in table (3).

The results of DPCA are promising for all data sets except data-set number 23 and 39. Operators reported no leak for cases 23 and 39, but DPCA resulted in leakage because the *SPE* index as shown in (23(c)) violated the threshold. These two problems need to be discussed with the company and corresponding refinements (adaptation rules) will be applied. The results for DPCA and PCA are summarized in the Table 3.

References

- [1] B. Rashidi and Q. Zhao, "Phase one progress report," UG Reports, 2016-Oct.
- [2] C. Alcala and S. Qin, "Reconstruction-based contribution for process monitoring," *Automatica*, vol. 45, no. 7, pp. 1593 1600, 2009.
- [3] A. Webb, "Statistical pattern recognistion," Wiley, vol. 2, 2003.

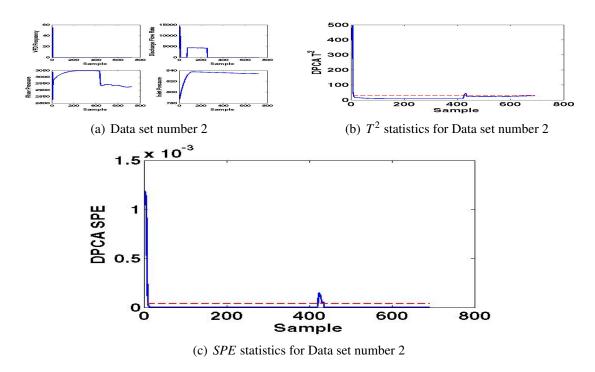


Figure 3. T^2 and SPE indices calculated by DPCA for Data set number 2

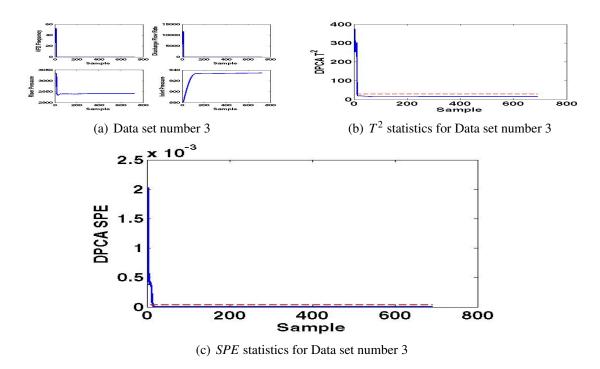


Figure 4. T^2 and SPE indices calculated by DPCA for Data set number 3

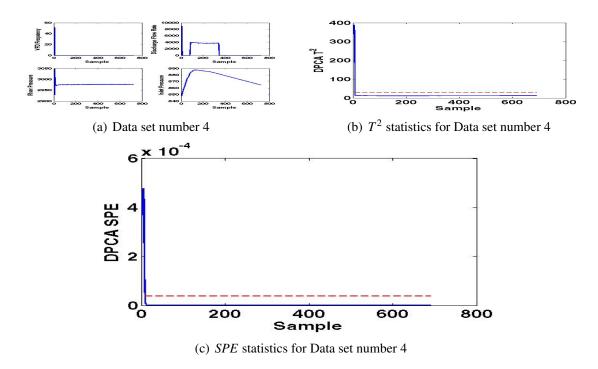


Figure 5. T^2 and SPE indices calculated by DPCA for Data set number 4

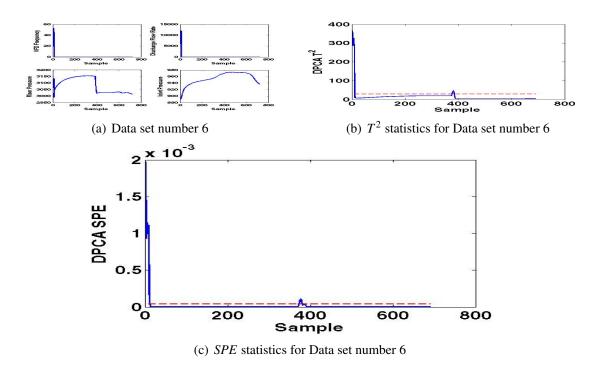


Figure 6. T^2 and SPE indices calculated by DPCA for Data set number 6

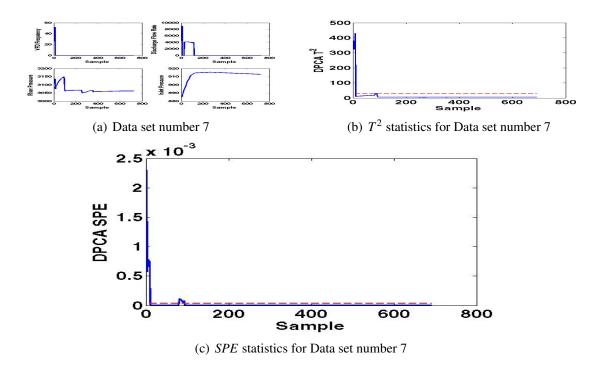


Figure 7. T^2 and SPE indices calculated by DPCA for Data set number 7

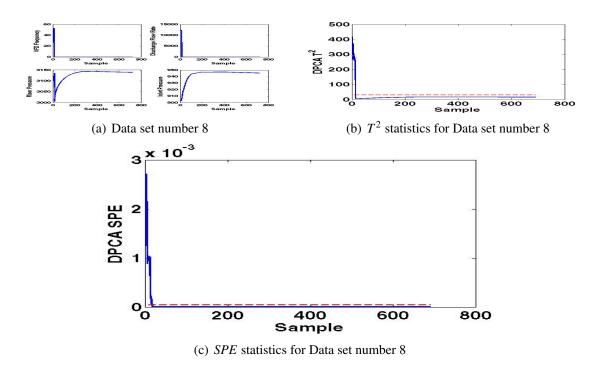


Figure 8. T^2 and SPE indices calculated by DPCA for Data set number 8

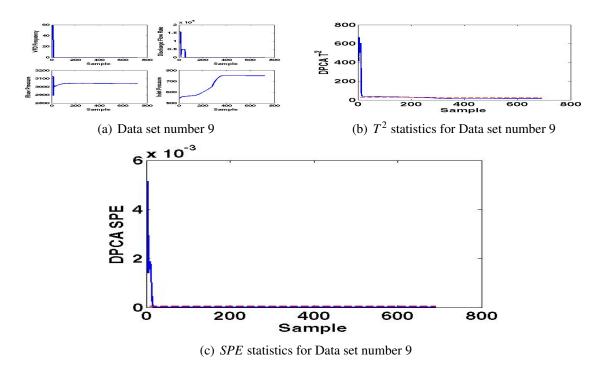


Figure 9. T^2 and SPE indices calculated by DPCA for Data set number 9

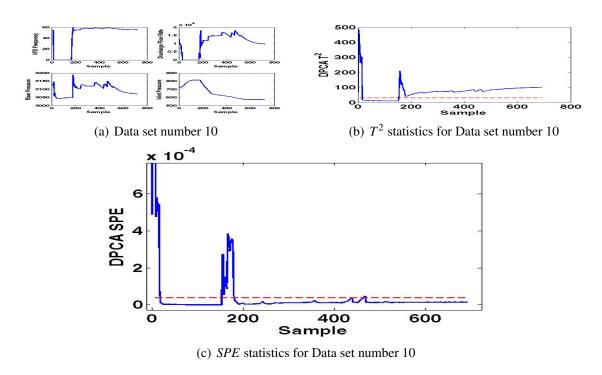


Figure 10. T^2 and SPE indices calculated by DPCA for Data set number 10

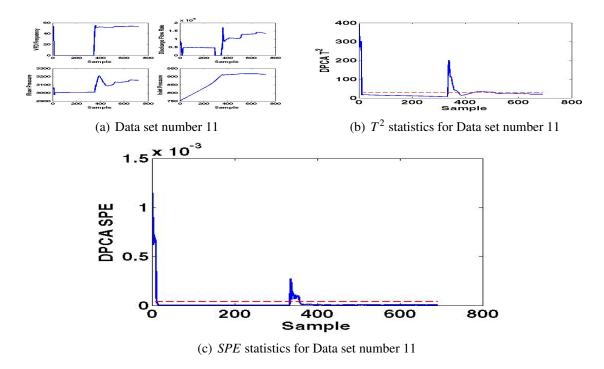


Figure 11. T^2 and SPE indices calculated by DPCA for Data set number 11

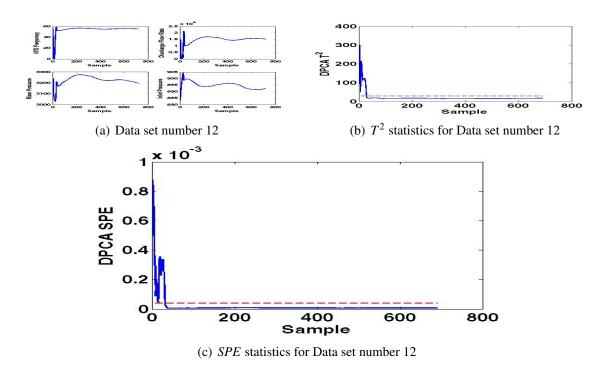


Figure 12. T^2 and SPE indices calculated by DPCA for Data set number 12

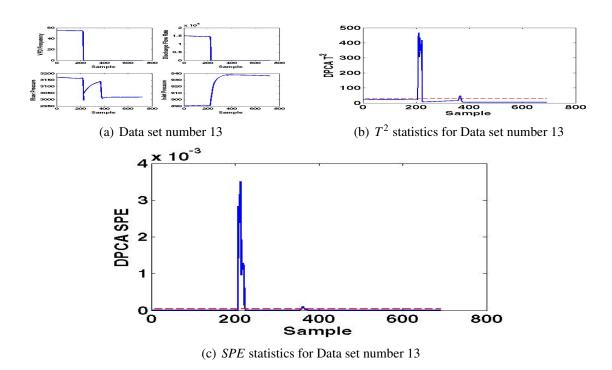


Figure 13. T^2 and SPE indices calculated by DPCA for Data set number 13

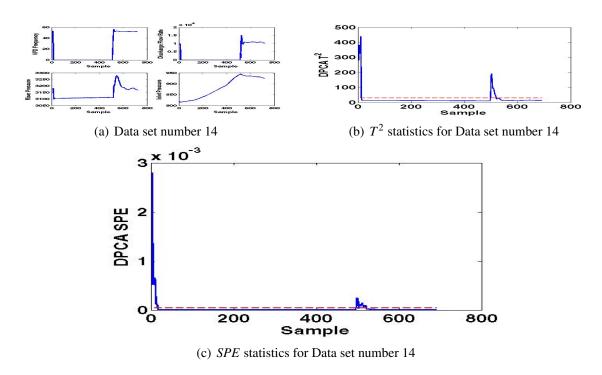


Figure 14. T^2 and SPE indices calculated by DPCA for Data set number 14

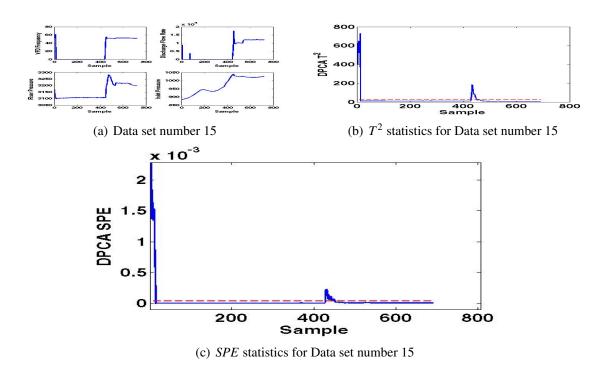


Figure 15. T^2 and SPE indices calculated by DPCA for Data set number 15

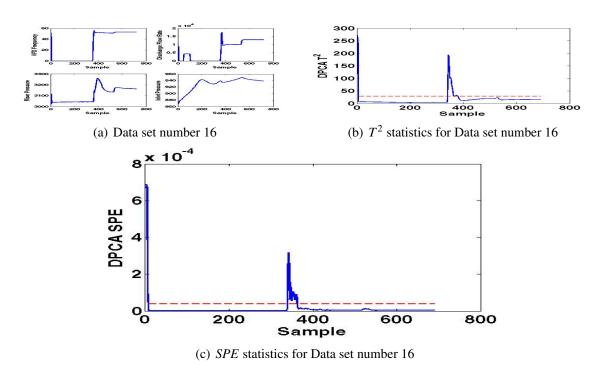


Figure 16. T^2 and SPE indices calculated by DPCA for Data set number 16

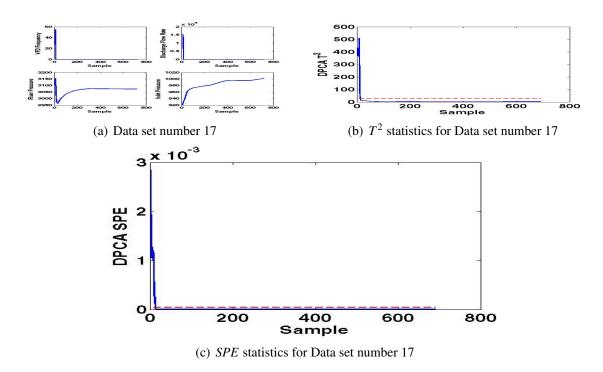


Figure 17. T^2 and SPE indices calculated by DPCA for Data set number 17

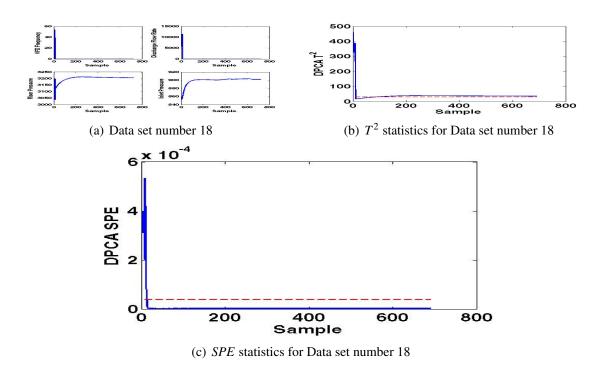


Figure 18. T^2 and SPE indices calculated by DPCA for Data set number 18

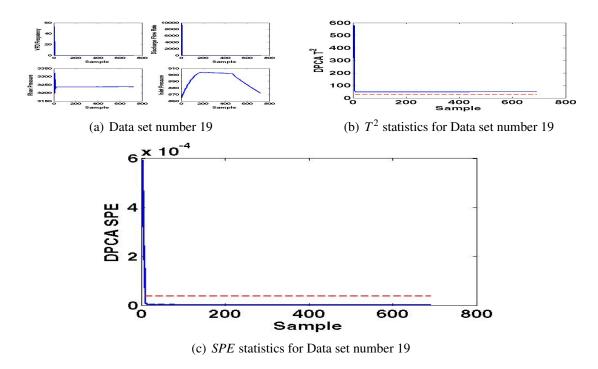


Figure 19. T^2 and SPE indices calculated by DPCA for Data set number 19

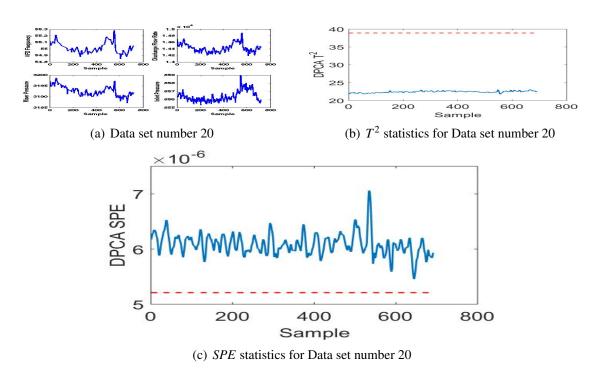


Figure 20. T^2 and SPE indices calculated by DPCA for Data set number 20

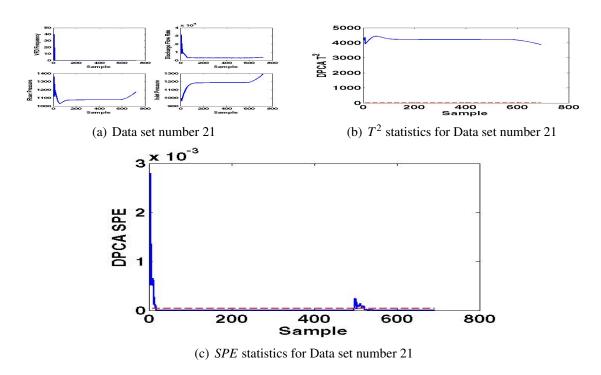


Figure 21. T^2 and SPE indices calculated by DPCA for Data set number 21

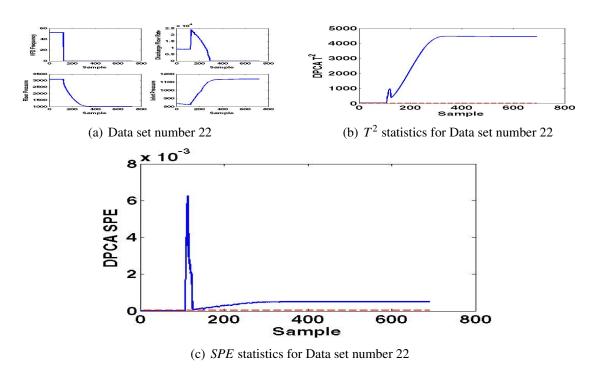


Figure 22. T^2 and SPE indices calculated by DPCA for Data set number 22

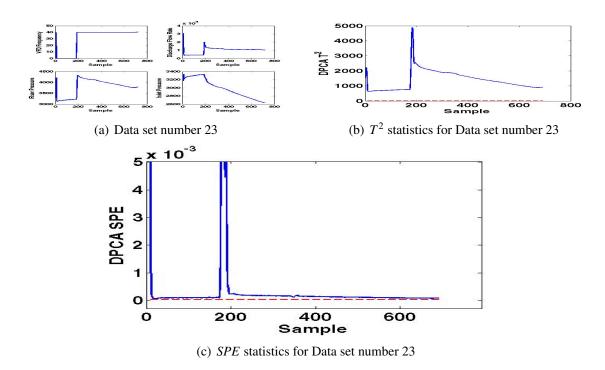


Figure 23. T^2 and SPE indices calculated by DPCA for Data set number 23

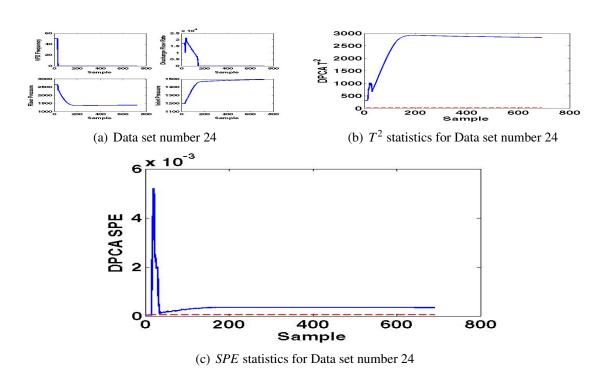


Figure 24. T^2 and SPE indices calculated by DPCA for Data set number 24

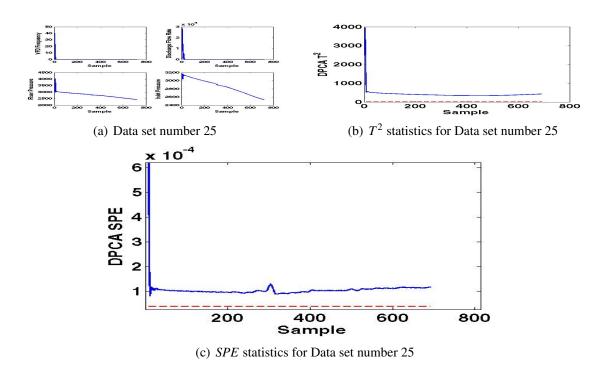


Figure 25. T^2 and SPE indices calculated by DPCA for Data set number 25

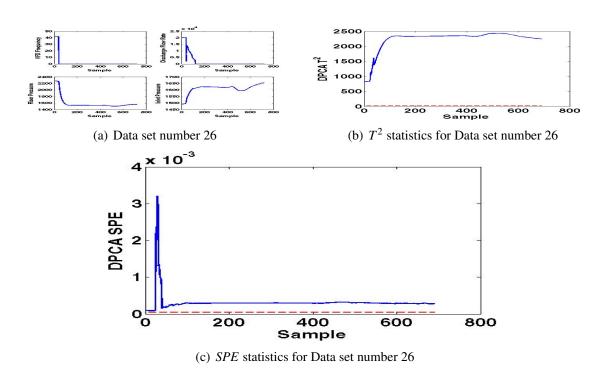


Figure 26. T^2 and SPE indices calculated by DPCA for Data set number 26

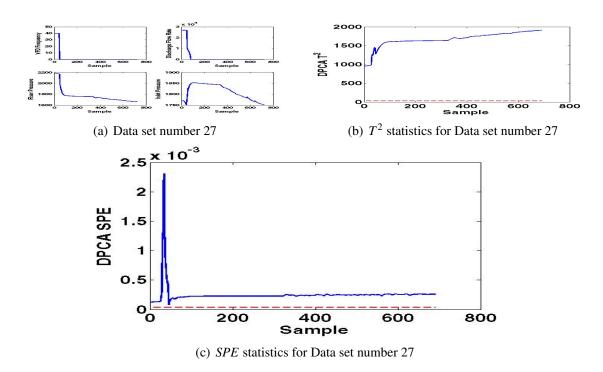


Figure 27. T^2 and SPE indices calculated by DPCA for Data set number 27

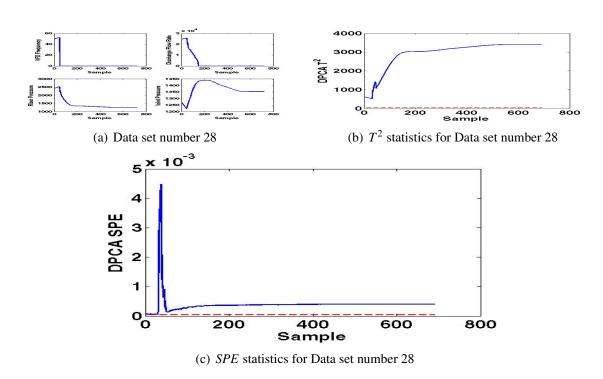


Figure 28. T^2 and SPE indices calculated by DPCA for Data set number 28

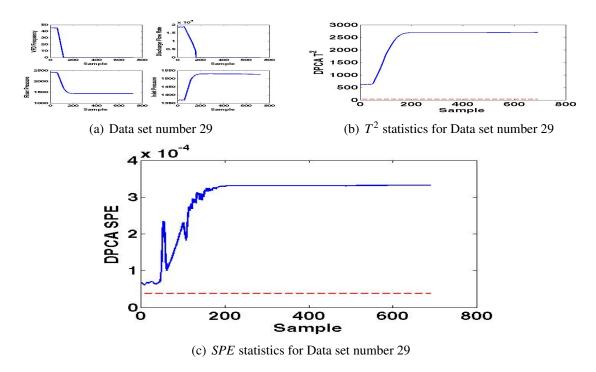


Figure 29. T^2 and SPE indices calculated by DPCA for Data set number 29

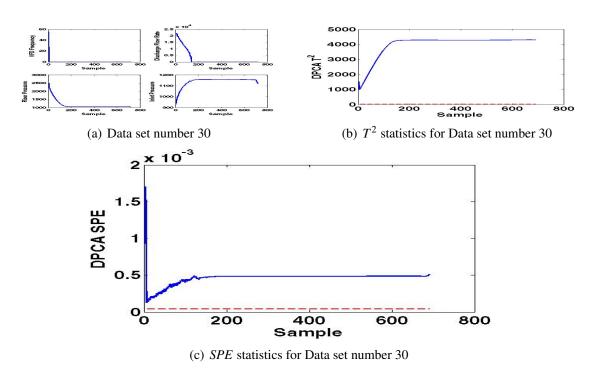


Figure 30. T^2 and SPE indices calculated by DPCA for Data set number 30

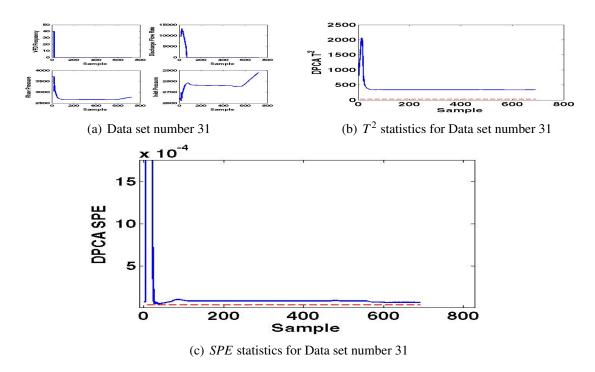


Figure 31. T^2 and SPE indices calculated by DPCA for Data set number 31

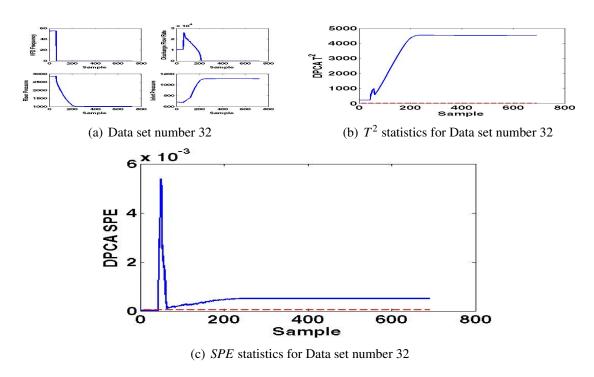


Figure 32. T^2 and SPE indices calculated by DPCA for Data set number 32

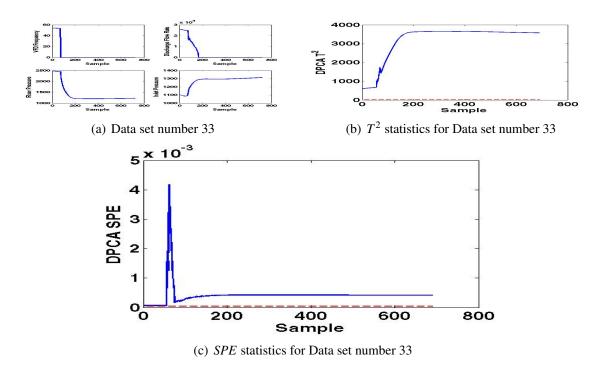


Figure 33. T^2 and SPE indices calculated by DPCA for Data set number 33

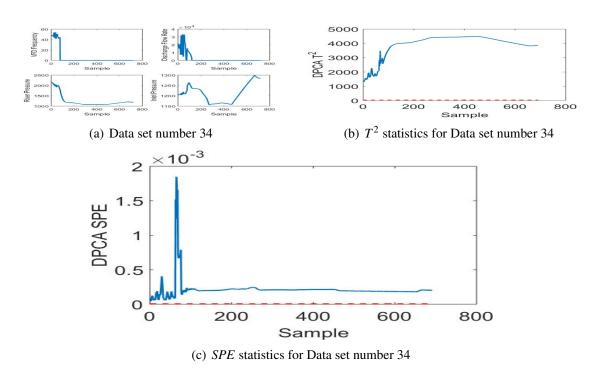


Figure 34. T^2 and SPE indices calculated by DPCA for Data set number 34

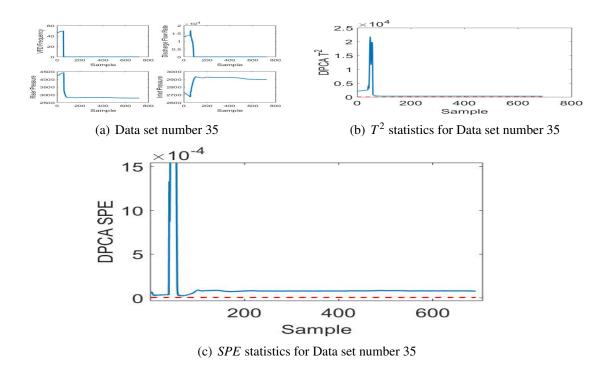


Figure 35. T^2 and SPE indices calculated by DPCA for Data set number 35

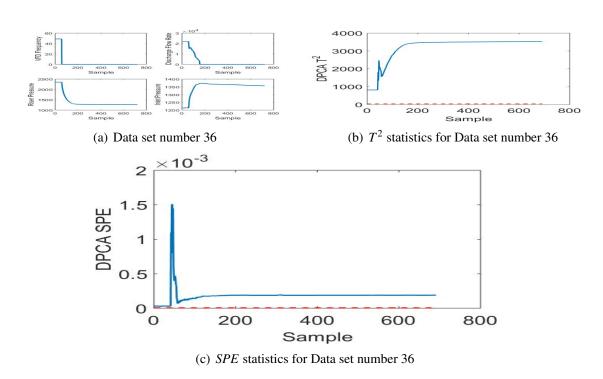


Figure 36. T^2 and SPE indices calculated by DPCA for Data set number 36

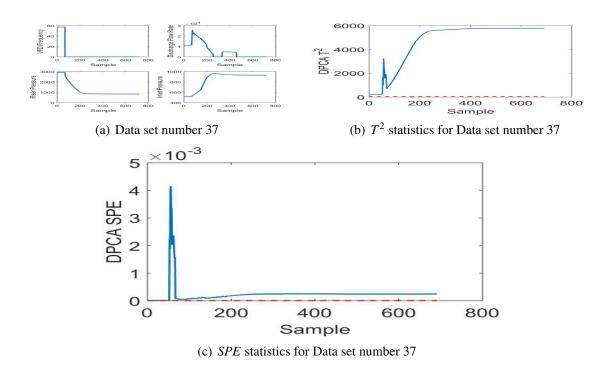


Figure 37. T^2 and SPE indices calculated by DPCA for Data set number 37

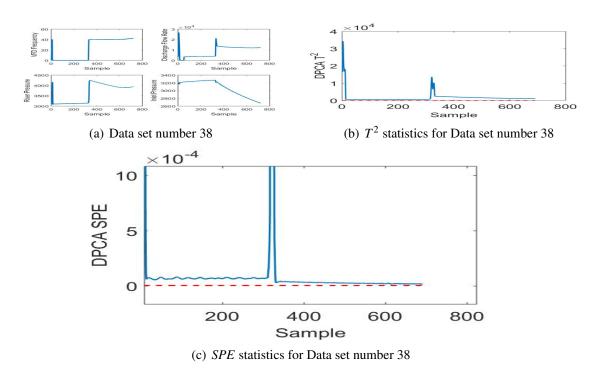


Figure 38. T^2 and SPE indices calculated by DPCA for Data set number 38

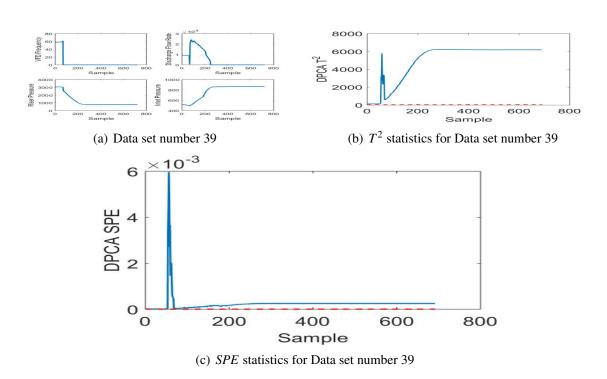


Figure 39. T^2 and SPE indices calculated by DPCA for Data set number 39

Table 1. Comparison Results for anomalies detection using DPCA and operators inspection

Data set number	Operators' inspection	DPCA	PCA
Data-set number 2	No leak	No leak	No leak
Data-set number 3	No leak	No leak	No leak
Data-set number 4	No leak	No leak	leak
Data-set number 5	No leak	No leak	No leak
Data-set number 6	No leak	No leak	leak
Data-set number 7	No leak	No leak	No leak
Data-set number 8	No leak	No leak	leak
Data-set number 9	No leak	No leak	No leak
Data-set number 10	No leak	No leak	No leak
Data-set number 11	No leak	No leak	leak
Data-set number 12	No leak	No leak	No leak
Data-set number 13	No leak	No leak	No leak
Data-set number 14	No leak	No leak	No leak
Data-set number 15	No leak	No leak	No leak
Data-set number 16	No leak	No leak	leak
Data-set number 17	No leak	No leak	No leak
Data-set number 18	No leak	No leak	leak
Data-set number 19	No leak	No leak	leak
Data-set number 20	leak	leak	No leak
Data-set number 21	No leak	No leak	leak
Data-set number 22	leak	leak	leak
Data-set number 23	No leak	leak (low leak)	leak
Data-set number 24	leak	leak	leak
Data-set number 25	low leak	leak	leak
Data-set number 26	leak	leak	leak
Data-set number 27	leak	leak	leak
Data-set number 28	leak	leak	leak
Data-set number 29	leak	leak	leak
Data-set number 30	leak	leak	leak
Data-set number 31	leak	leak	leak
Data-set number 32	leak	leak	leak
Data-set number 33	leak	leak	leak
Data-set number 34	leak	leak	leak
Data-set number 35	leak	leak	leak
Data-set number 36	leak	leak	leak
Data-set number 37	leak	leak	leak
Data-set number 38	leak	leak	leak
Data-set number 39	No leak	leak	leak