What Went Wrong and When?

Multivariate Fault Diagnosis of a High Purity Distillation Column

Isuru A. Udugama[1], David I. Wilson[2], Tajammal Munir[1], Brent Young[1] and Wei Yu[1,*]

[1]Chemical & Materials Eng. Dept., The University of Auckland, New Zealand
[2]School of Engineering, Auckland University of Technology, New Zealand
*w.yu@auckland.ac.nz

Abstract: Distillation is a widely used industrial separation process. Process monitoring using multivariate data analysis for such capital intensive processes has the potential to reduce energy and resource requirements. Notwithstanding the large body of work in industrial process monitoring, little prior work has focused on monitoring ultra-high purity multicomponent distillation processes. This extreme situation significantly increases the difficulties, particularly when considering the underlying nonlinearities.

In this paper, principal component analysis (PCA) was employed to analyze the process variables in order to obtain a better understanding of the key relationships affecting the product quality. Using this data-driven approach, it was found that the reboiler steam flow rates were the most important factor affecting the top product specification. These results were subsequently validated by the plant engineers who attributed the abnormal variation in the steam supply system to other unit operations. It provides a potential solution to guide the control system to maintain a narrow variation in the top product ppm in order to meet specification.

Keywords: High purity distillation column, Principal component analysis.

1. INTRODUCTION

Distillation is a widely used industrial separation process. The control and optimization of distillation processes has been the focus of many academic studies as potential improvements have a real impact on reducing energy and resource usage of industrial process. Despite there being a large body of work on this general area no previous work has focused on monitoring industrial distillation processes where impurities entering at ppm (parts per million) levels need to be further separated out of the final product to arrive at ultra-high purity products. This is due to the relative complexity of monitoring and controlling products at high purity and the nonlinearities associated with operating at this point.

Distillation is the dominant process for separating large multi-component streams into high purity products. In the refining and chemical industries it consumes approximately 40% of the total energy used for operating the plant which is approximately 3% of global energy consumption (Annakou and Mizsey, 1996, Englinen et al., 2003). Optimization is an efficient way to improve the energy efficiency and product quality.

Optimization of distillation columns is however not a straightforward process. Firstly, columns come in a variety of configurations with different operating objectives. These differences cause distinct dynamic behaviours and different operational degrees of freedom. Secondly, this situation becomes more complex when the product purity is regulated down to the 10~1000 parts-per-million (ppm) range due to the complex dynamics and non-linear behaviour. This highly nonlinear behaviour (e.g. as reported in Luyben et al., 1992 and Blevins et al., 2003) is a great challenge for control and optimization. Thirdly, column operations normally exhibit significant interactions within the control loops (i.e. composition control loops) and have numerous constraints. Lastly, disturbances from upstream or utilities may result in difficulties in obtaining and/or maintaining the optimized conditions. In this paper, we will study a distillation column which is used to produce a high purity methanol from methanol, ethanol, butanonl and water mixtures. Since variations from upstream and utility units i.e. low pressure steam, the ethanol concentration in the top product shows fluctuations which may violate the specification of the ethanol (ppm) upper limit. In this paper, we will use principal component analysis (PCA) to monitor the column operating conditions so that the ethanol specification can be secured with the optimal operating conditions.

2. PROCESS DESCRIPTION

For our study case, each distillation unit consists of a topping column and a refining column. The crude methanol, (which consists of methanol (80%), ethanol (150ppm w/w), butanol (95ppm w/w), water (20%) and some light gases), is first

Preprint submitted to 9th International Symposium on Advanced Control of Chemical Processes. Received November 18, 2014.

pumped into the topping column where the light gases are taken off in the top stream. The bottoms stream is sent through a series of condensers and heat exchangers and then fed into a refining column (shown in Fig. 1) with a large number of stages, 87 in this case study. Methanol (product) is taken from the top of the refining column with 99.99% purity, while water is taken from the bottom. Ethanol is predominantly taken out in the side- draw (it also called as a fusel stream), located near the middle of the column. The design operating condition is listed in table 1. The product flow rate is mainly governed by the concentration of ethanol (ppm) in the product stream, where a strict level is enforced (e.g. 10.07 ppm in table 1). Failing to meet the product ethanol specification would result in a substandard product with a lower commercial value.

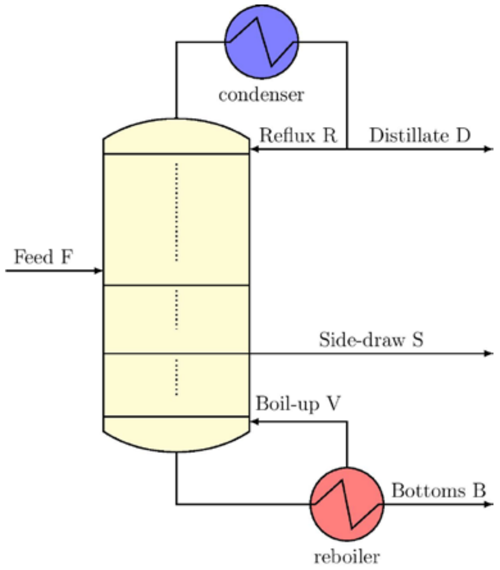

Fig. 1: Schematic of high purity distillation column

Table 1: Distillation operating conditions

location			Design
Feed (F)	Flow rate	kg/hr	107430
	Methanol	Mass fraction	0.79
	Ethanol	ppm (mass fraction)	150
	Water	mass fraction	0.21
	Butanol	Ppm	600
Top Product (D)	Flow rate	kg/hr	83333
	Methanol	mass fraction	0.99996
	Ethanol	Ppm	10.07
	Water	Ppm	56.23
	Butanol	Ppm	0
Bottom Product (B)	Flow rate	kg/hr	22026
	Methanol	Ppm	17.79
	Ethanol	Ppm	0.0
	Water	mass fraction	1.0
	Butanol	mass fraction	0.0
Fusel (S)	Flow rate	kg/hr	2031
	Methanol	mass fraction	0.53
	Ethanol	mass fraction	0.05
	Water	mass fraction	0.34
	Butanol	Ppm	6000
Reflux ratio			2.28
Reboiler duty (MW)			Unknown
Methanol recovery			0.987

3. PRINCIPAL COMPONENT ANALYSIS

3.1 Basics

Principal Component Analysis (PCA) is a traditional multivariate statistical technique commonly applied for the linear conversion of a multivariable space into a new subspace while retaining maximum variance from the original space with the least number of dimensions possible. PCA is widely used to analyse multivariate data for data reduction. Extensive literature is available on the procedure and applications of this technique (e.g. Kourti and MacGregor 1995; Jolliffe 2002; Saporta and Niang 2009).

The general procedures are described here briefly:
a) The matrix $\mathbf{X}_{n*m}$ contains the measured data where each row is the simultaneous observation of the m measured variables. The aim of PCA is to find the subset of the P principal components which retain most of the information that is in the original data set.
b) $\mathbf{X}_{n*m}$ is normalized to have zero mean (based on columns).
c) The covariance matrix **R** is constructed as follows:

$$\boldsymbol{R} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} \qquad (1)$$

d) A singular value decomposition (SVD) is performed on **R** as follows

$$\mathbf{R}=\mathbf{V\Lambda V}^{\mathrm{T}} \qquad (2)$$

where $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues of **R** in decreasing numerical order, and **V** is the matrix of the eigenvectors of **R**.
e) Matrix **P** is constructed by selecting the first a columns of **V** (eigenvectors) corresponding to the first a principal eigenvalues.
Several procedures have been proposed for the selection of the number of principal components (a) to retain as much significant information as possible, such as Cumulative Percentage Variance (percentage variance retained by the first "a" principle components) and cross validation.
f) Matrix **P** is used to transform the original space of variables to the reduced dimension subspace:

$$\mathbf{T} = \mathbf{XP} \qquad (3)$$

The columns of matrix $\boldsymbol{P}$ are the selected eigenvectors and are called loadings. The elements of matrix **T** are the values of the original measured variables which have been transformed to a reduced subspace and are called scores.

3.2 PCA-based process monitoring

PCA has become one of the most popular multivariate statistical methods for monitoring and fault detection in many different fields, e.g. aluminium processing (Majid et al., 2012), oil refining (Chen et al., 2004; Kourti and MacGregor, 1995), aircraft structure (Mujica et al., 2010) and social science (Dunteman, 1989).
PCA is one of the most effective tools for detecting abnormal behaviours (or faults) in a process or system. Hotelling's T^2 and Q statistics are commonly used for PCA fault detection. Hotelling's T^2 statistic is based on

Preprint submitted to 9th International Symposium on Advanced Control of Chemical Processes. Received November 18, 2014.

analysing the score matrix **T** to check the degree to which the data fits the calibration model. Hotelling's T^2 value for the i-th observation is,

$$T_i^2 = t_i^T \Lambda^{-1} t_i \quad (4)$$

where $\Lambda = diag(\lambda_1, \ldots, \lambda_k)$ is a diagonal matrix of PC eigenvalues, and $\mathbf{t}_i$ is the PC score for the i-th observation. An Upper Control Limit (UCL) for Hotelling's T^2 is given by the following equation:

$$T_{UCL}^2 = \frac{(n^2-1)k}{n(n-k)} F_\alpha(k, n-k) \quad (5)$$

where $F_\alpha(k,n\text{-}k)$ is the upper $100\times(1\text{-}\alpha)\%$ confidence interval of the F distribution with k and $(n\text{-}k)$ degrees of freedom.

The Q-statistic is based on analysing the residual data to represent the changes in the data that are not explained by the PCA model. The Q-statistic of the i-th observation (sample), $\mathbf{x}_i$, is defined as follows:

$$Q_i = x_i(I - PP^T)x_i^T \quad (6)$$

where P matrix is the loadings matrix of PCA model. The $100\times(1\text{-}\alpha)\%$ control limit of the Q-statistic is given by Jackson and Mudholkar (1979) as follows:

$$Q_{UCL} = \theta_1 \left[\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_1 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \quad (7)$$

where $\theta_i = \sum_{j=k+1}^{p} \lambda_j^i,\ i = 1,2,3,\ h_0 = 1 - (2\theta_1\theta_3)/3\theta_2^2$ and z_α is the upper α percentage point of the standard normal distribution.

Generally, the Hotelling T^2 statistic exhibits large variance and therefore the abnormal behaviour can only be detected if there is great change or difference between the normal and the abnormal cases. The Q-statistic is much more sensitive than Hotelling's T^2 statistic since the Q_i values are relatively small compared to the T_i^2 values. The Q-statistics can therefore distinguish minor changes or differences.

In process monitoring and fault detection, when the process is operated as normal, the variation of the Hotelling's T^2 and Q statistics stay within the nominal region (control limits) specified by engineers. Based on reference data, a PCA model can be built although it should be noted that the reference data should be carefully selected so that it does not include any abnormal behaviour. It is also important that the reference data set should be large enough to reflect the normal situation.

If the abnormal observation is detected by the Hotelling's T^2 chart the diagnosis procedure can be carried out in following two steps:

i) normalized scores for that observation ($T_{i,a}/\lambda_a$) are calculated and the a-th score with the highest normalized value is selected.

ii) The contribution of each original k-th variable to this a-th score at this abnormal observation is given by,

$$\text{Contribution}(T_{i,a}) = X_{i,k} \times P_{a,k} \quad (8)$$

where $P_{a,k}$ is the loading of the k-th variable at the a-th component.

A bar chart of these contributions can be plotted, and variables with high contribution but with the same sign as that score will be investigated (contributions of the opposite sign will only make the score smaller).

4. DATA

Table 2 lists the 23 process variables collected at 3 minute intervals from an industrial high purity distillation column used for a PCA analysis The top product quality variable, ethanol concentration (in ppm) was also used to build the bridge between PCA analysis and quality analysis. The data matrix consists in total of n=25684 samples which correspond to just under around 2 months of operation.

Table 2: Process variable descriptions

Variable	Flow rate	Variable	Temp.	Variable	Temp.
1	Reboiler B steam	8	Tray 87	16	Tray 12
2	Product	9	Tray 84	17	Tray 10
3	Fusel	10	Tray 64	18	Tray 8
4	Bottom	11	Tray 44	19	Tray 6
5	Reboiler A steam	12	Tray 32	20	Tray 2
6	Reboiler duty (MW)	13	Tray 26	21	Tray 20
7	Column pressure	14	Tray 16	22	Tray 74
		15	Tray 14	23	Tray 54

5. BUILDING A SUITABLE PCA MODEL

To monitor this high purity distillation column, we need to build a PCA model for the normal operating conditions. Since methanol purity is the one of the most important controlled variables, ethanol mass fraction in the top product stream is selected as an indicator for selecting the normal operating period. The ethanol mass fractions were plotted in Fig. 2.

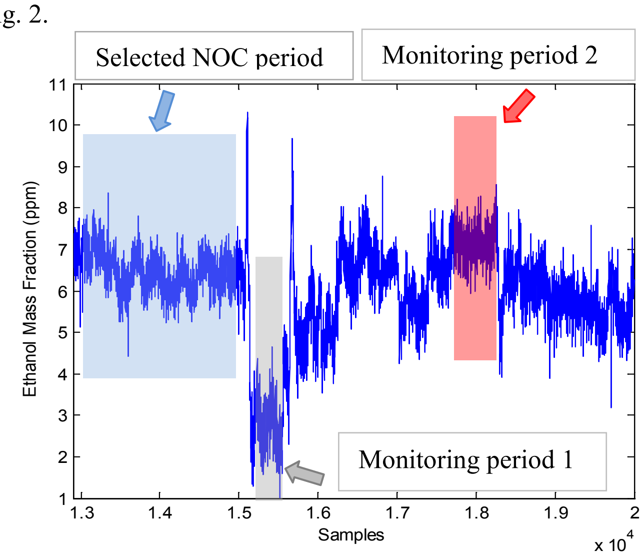

Fig. 2: Ethanol mass fractions in top product stream

From Fig. 2, sample periods from 12900 to 15000 is selected as the 'normal operating conditions (NOC)'. The average methanol mass fraction in this period is 6.51 ppm which is close to the normal operating set point. The variance in this period is 0.237 which is statistically significantly smaller than the whole time series variance: 2.05.

Since the variables in the **X** matrix contain variables of different magnitude (e.g. mass and temperature values), the data are scaled to unit variance. The first two principal components from the PCA model built from the NOC period are shown in Fig. 3 (the black ellipse shows the 95% confidence-limit).

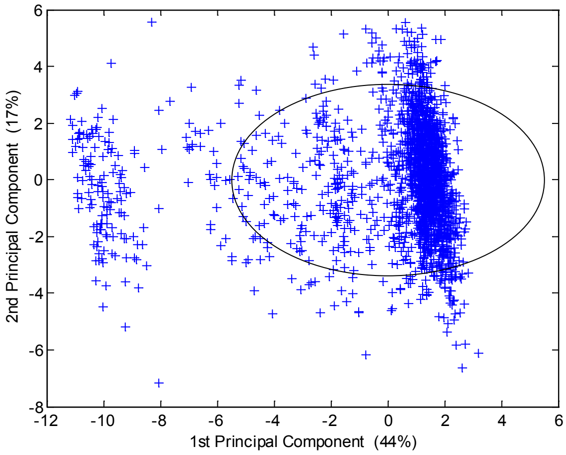

Fig. 3: Score plot of the first two PCs

A Pareto chart with the percentage of variance explained by each principal component is plotted in Fig. 4. It shows that 83% of the total variance can be represented by the first 6 principal components.

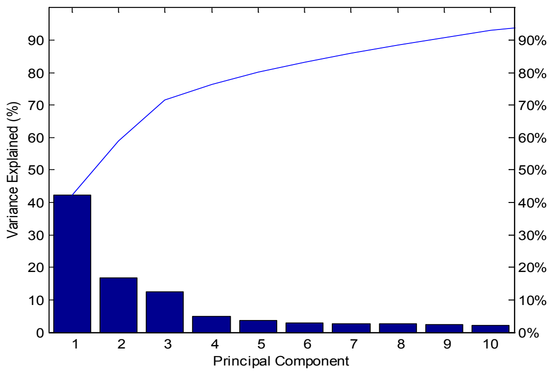

Fig. 4: Variance explained by the principal components

6. PCA MONITORING

6.1 Detecting a significant operating change

Fig. 2 shows the ethanol values dropping rapidly after the selected NOC period (denoted 'Monitoring period 1' in Fig. 2) and this period will be monitored by a PCA model regressed during the initial NOC period. The first 3 principal components for the samples period from 15001 to 16500 are plotted in Fig. 5. We can see two clusters corresponding to the 2 operating regions and a short transition region between them. The first cluster about the origin and corresponds to the NOC period. The second cluster corresponding to the operating period around [15001 15500] where the average ethanol mass fraction is 3.9 (ppm). This can potentially be a time period where the plant is running at different rates/recovery due to disturbance.
The contribution plot for a point in the second cluster is illustrated in the bottom subplot in Fig. 6. The top subplot in Fig. 6 shows the contribution plot from one point in the NOC period. It shows that the variable 2 and 4 (corresponding to Top and Bottom flowrates respectively) are the two dominant variables for these periods of operation. The reboiler duty is relatively stable at this time period, this in turn should make the energy balance stable, which means any variations in ethanol compositions not due to reboiler duty. However, the product ethanol composition is also influenced by recovery and mass balance. In this instance variations in variable 2 (product flow rate) will contribute to the ethanol variation. Variation in the product flow rate will then influence the bottoms flow rate (Variable 4) due to the mass balance. These interactions are captured in Fig. 6 (top graph).

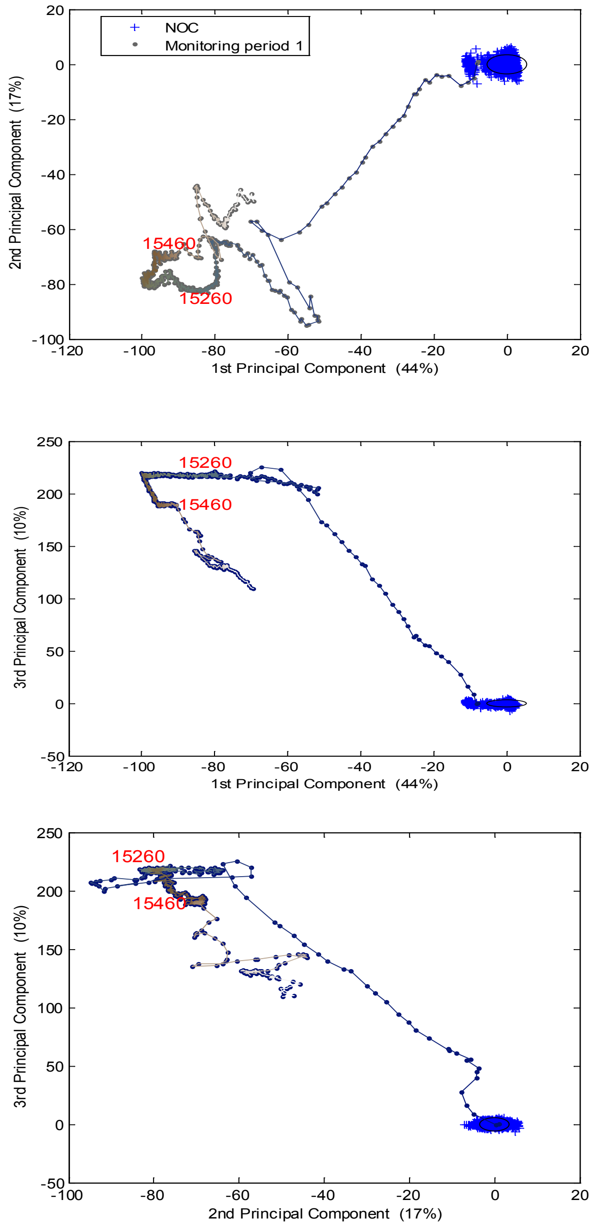

Fig. 5: First 3 PC score plots

In contrast the main contribution variable during the abnormal observation period is variable 5 (reboiler A steam flowrate). By looking at variable 5 during that period in Fig. 7, we observed that the flow rate dropped from 63 ton/hr to

35 ton/hr. In this instance the energy balance has been significantly violated, to remedy this issue, the mass balance should be brought to a new feasible equilibrium. This has been achieved by reducing the product flow rate as well as the feed rate as seen in Fig. 7.

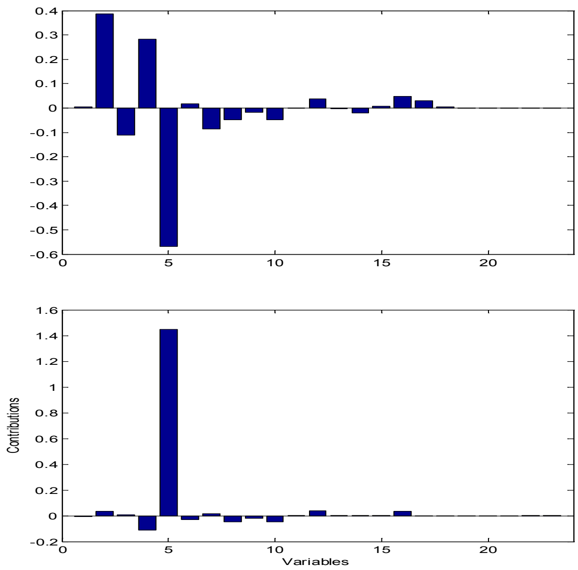

Fig. 6: Contribution plots for the selected components

The product flow rates for the corresponding observed period were also plotted in Figure 7. We can observe a significant product flow rate drop which may be the main reason to drive the process away from the NOC. Unfortunately, the feed flow rate was not recorded in this data set, but the bottom flow rate did not change too much. From this information, we may assume that the feed was also significantly reduced from sample 15060. This may be the true reason for the process drifting away from the NOC.

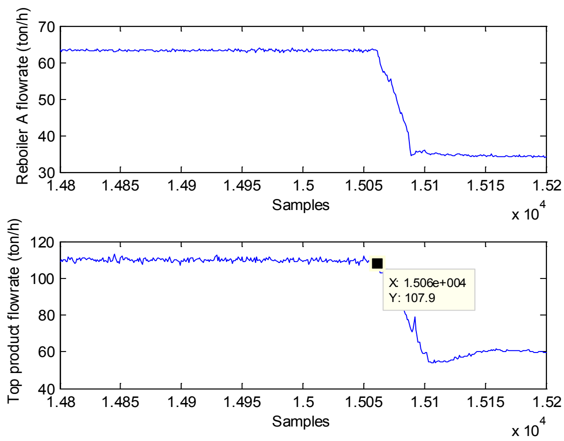

Fig. 7: Reboiler A flow rate and Top product flow rate

The Hotelling's T^2 plot in Fig. 8 clearly indicated that the process was drifting away from the defined NOC region. These drifts did significantly affect the top product ethanol mass fractions to reach an average 3.9 ppm for 'Monitoring period 1' shown in Fig. 2.

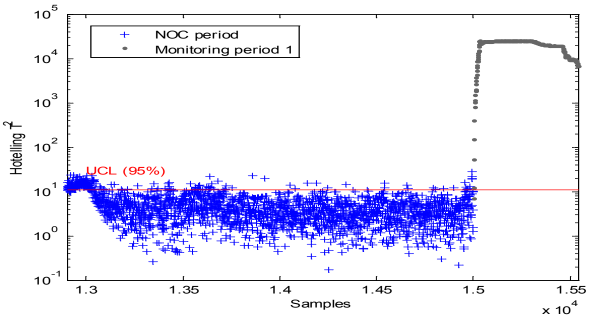

Fig. 8: Hotelling's T^2 plot

6.2 Minor process changes

The plant deviation analysed in Section 6.1 is significant so it is not surprising that the PCA analysis detected an abnormal condition. This section however is concerned with the 'Monitoring Period 2' given in Fig. 3 where the process has apparently recovered back to the normal operating condition if we restrict our observation only to the ethanol concentration. This is a much stronger test for the proposed process monitoring analysis since it must uncover much more subtle and latent changes in the underlying variables. The first 3 PC score plots for the sample period from 17800 to 18100 (denoted 'Monitoring period 2' in Fig. 2) are shown in Fig. 9.

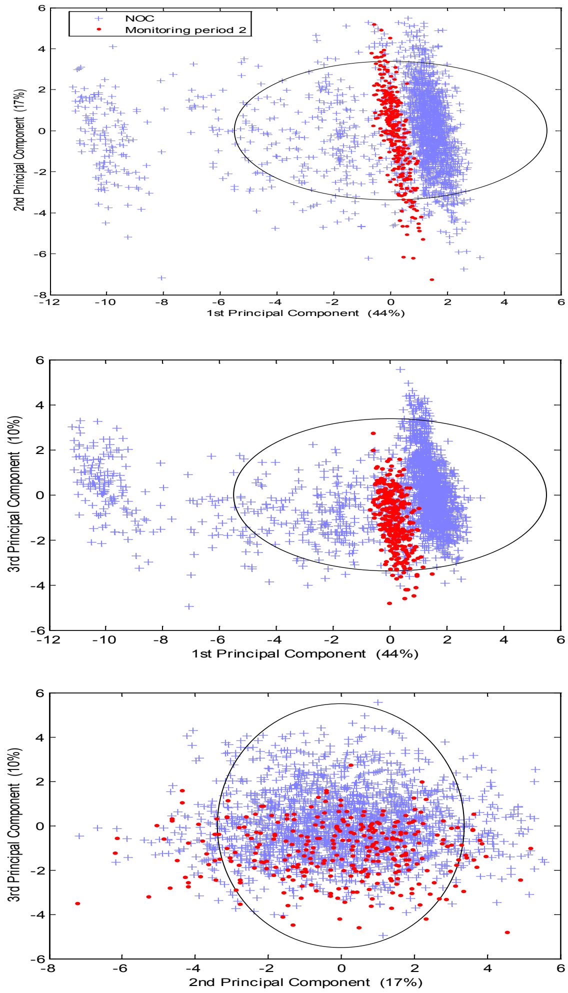

Fig. 9: The first 3 PC score plots for 'Monitoring period 2'

From Fig. 9, we observe that the process is controlled within the NOC. The Hotelling's T^2 and Q statistic plots shown in Figure 10 also support this observation.

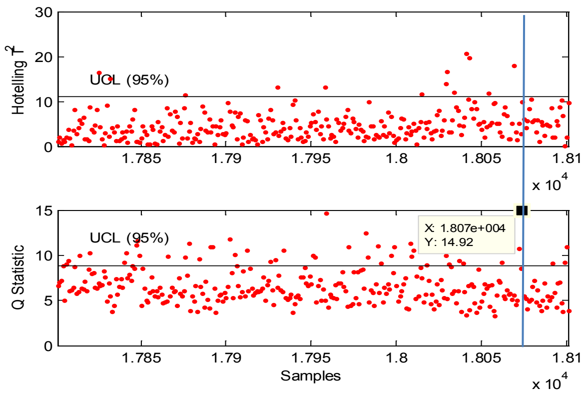

Fig. 10: Hotelling's T^2 and Q-statistic plots for 'Monitoring period 2'

The Q statistic plot, being more sensitive than the Hotelling's T^2, does however show that the process is operating slightly away from the NOC. The contribution plot for the observations during sample period 18070 is illustrated in Fig. 11.

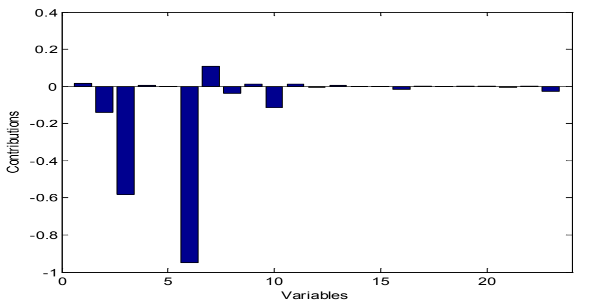

Fig. 11: Contribution plot for a selected observation time (sample period 18070) in 'Monitoring period 2'

Variables 3 and 6 corresponding to fusel flowrate and Reboiler duty respectively are the dominant variables that affect this observation time. The fusel flowrate and reboiler duty process variables for that period are plotted in Fig. 12.

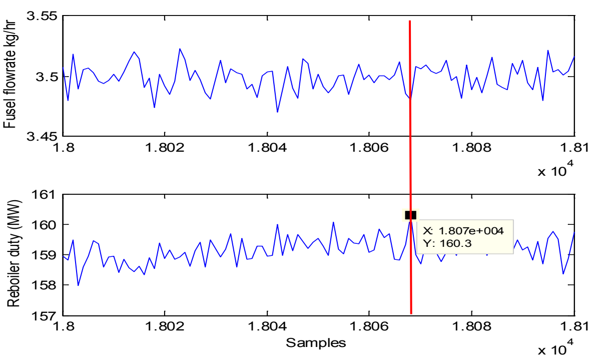

Fig. 12: Fusel flowrate (top) and Reboiler duty (bottom)

This shows a spike at sample 18070 which directly drives the process away from the NOC. As mentioned before any change in reboiler duty affects the energy balance. To understand the effect of fusel flow rate, one needs to think of the ethanol species balance with in the column. Despite being a small stream almost 95% of the ethanol entering the column needs to leave at the fusel draw point. As such even a slight change in the fusel draw can potentially create issues in the short run.

7. CONCLUSIONS

In this paper, PCA was implemented on a high purity distillation column for process monitoring purposes. The PCA analysis significantly contributed to understanding different periods of operations and dominating variables for these periods. PCA showed that the reboiler duty plays the most important role for the top product ethanol specification control. The Hotelling's T^2 chart clearly indicates the process drifting from the defined NOC when process excursions are significant. However when the excursions are less severe, or when the process recovers apparently to the normal operating condition, the more sensitive Q statistic plot could detect the small drift from the NOC. Further variable contribution analysis may help us to identify the reason causing the drift.

REFERENCES

Blevins, T.L., McMillan, G.K., Wojsznis, W.K., Brown, M.W. (2003). Advanced Control Unleashed: Plant Performance Management for Optimum Benefit, *International Society for Automation (ISA)*, Research Triangle Park, NC.

Chen, Q., Kruger, U., Meronk, M. and Leung, A.Y.T. (2004). "Synthesis of T^2 and Q statistics for process monitoring." *Control Engineering Pratice* **12**: 745-755.

Dunteman, G.H. (1989). *Principal Component Analysis*. London, SAGE Publicaitons Inc.

Engelien, H.K., Larsson, T., Skogestad, S. (2003). Implemenation of optimal operation for heat integrated distillation columns. *Chemical Engineering Research and Design*, 81(2): 277-281.

Jackson, J. E. and Mudholkar, G. (1979). "Control procedures for residuals associated with principal component analysis." *Technometrics* **21**: 341-349.

Jolliffe, I.T. (2002). *Principal component analysis*. Springer, New York.

Kourti, T. and Macgregor, J.F. (1995). Process analysis, monitoring and diagnosis, using multivartiate projection methods. *Chemometrics and Intelligent Laboratory Systems*, 28:3-21.

Luyben, W.L., Tyreus, B.D., Luyben, M.L. (1992). *Practical distillation control*. New York, Van Nostrand Reinhold.

Majid, N.A.A., Taylor, M.P., Chen, J.J.J., Yu, W. and Young, B.R. (2012). Diagnosing faults in aluminium processing by using multivariate statistical approaches. *Journal of Materials Science* **47**: 1268-1279.

Mujica, L.E., Rodellar, J., Fernandez, A., and Guemes, A. (2010). Q-statistic and T^2-statistic PCA-based measures for damage assessment in structures." *Structural Health Monitoring* **10**(5): 539-553.

Saporta, G. and Niang, N. (2009). Principal Component Analysis: Application to Statistical Process Control, in Data Analysis (ed G. Govaert), ISTE, London, UK.