공학박사학위논문

Fault Detection and Root Causality Analysis using Multi-mode PCA and Multivariate Granger Causality

다중 모드 주성분 분석과 다변수 그레인저 인과관계 기법을 이용한 이상 감지와 근본 원인 분석 방법론

2019년 8월

서울대학교 대학원

화학생물공학부

편 하 형

Abstract

Fault Detection and Root Causality Analysis using Multi-mode PCA and Multivariate Granger Causality

Hahyung Pyun

School of Chemical & Biological Engineering

The Graduate School of Seoul National University

Process data analysis has been great developed for decades in accordance with progress of data storage and processing speed. As a result, most of plant are not only using univariate methods, but also multivariate statistical methodologies for real time monitoring. From the analysis of accumulated normal data, detection accuracy and rate have been progressing.

However, unlike the fault detection area, fault diagnosis has many problem. Process fault diagnosis method is largely classified into three methodologies;

qualitative model based analysis, knowledge based analysis and historic data based analysis. From qualitative analysis perspective, as the process becomes large and more complicated, it it practically impossible to provide appropriate information for all abnormal situation. In terms of knowledge based analysis, such as expert system, the accuracy can be high, but it takes a long time to analyze the fault, so this method is generally used for post-accident diagnosis. Because of the limitations, methodologies for real time monitoring and diagnosis are mainly based on historical data. However, most algorithms of historical analysis use specific data driven model that actual data occurred in the past so it is only for used in that case.

To solve this problems of real-time fault diagnosis, this thesis proposes root cause analysis with the fault detected time simultaneously. Especially, it is focused on providing an accurate root causality even when the fault has a small intensity.

First, for the fast detection of abnormal situation, principal component analysis (PCA) method is used. Several methods integrated with PCA in normal operation data modeling procedures. T-score, derived from global PCA is classified into k normal mode. Divided normal operation data, local PCA models are developed respectively.

Second, minimum distance to mean (MDM) and k-nearest neighbors (kNN) are used for matching the class new samples with training normal data. And then, process is monitored by local mode in detail. When the fault is detected, with integration PCA contribution and singular value decomposition, hierarchy sensors are selected. From these sensors, MVGC analyzes root causality.

To verify performance of the proposed method, liquefied natural gas (LNG) plant fractionation dynamic model is used. From this dynamic model, 45 fault cases are simulated. Proposed method is perfectly better performance than global PCA. In terms of fault detection accuracy (FDA) and fault detection rate (FDR), 43 out of 45 cases show dramatically increased results and 2 cases the same results. Comparing with univariate, shewart 3-sigma, 35 cases are increased results, 8 cases same results and only 2 cases very lightly poor results. From the MVGC analysis, root cause analysis is compared with conventional contribution chart and residual subspace (RS) amplification. As a result, proposed method provides appropriate root cause while conventional contribution and RS amplification are failed to find root cause. Specially, root cause of developed method is similar with real time alarm later. This methodology provides root cause information only based on normal data, also suitable for small intensity fault, it is applicable to most process. It is expected to contribute greatly analysing the new fault in real time.

Keywords: Process monitoring, Fault detection and diagnosis, Multi-mode operation, Granger causality, Principal component analysis, Fault magnitude

Student ID: 2011-30989

Contents

List of Figures

List of Tables

CHAPTER 1 Introduction

1.1 Research motivation

Process monitoring has became an essential part of plant management. Before the computer science is developed, the main methodologies of monitoring and fault detection and diagnosis (FDD) are constructed qualitatively model based on knowledge and expert system. After the computer science and data processing are developed, data driven analysis have been developed sharply, and recently most studies are based on data driven. Some of them, knowledge and model based methodologies are integrated with data driven algorithm.

In data driven methodologies, generally, the task of process monitoring is consist of 4 parts.[1] ; fault detection, fault identification (or diagnosis), fault estimation and fault reconstruction. In fault detection part, there have been a great development of methodology due to the enormous data. Especially, principal component analysis (PCA), partial least square (PLS), fisher discriminant analysis (FDA) such as dimension reduction methods are used widely.[1] Although PCA method has mainly strength in a high dimension convert into a low dimension, it is also descending accuracy use enormous data and variables directly. Besides, this method can cover the linear relationship between variables, it is needed of other methodology for increasing accuracy. For this reason, reduction methodologies or low dimensioned variables can be adjusted according to purpose of data handling. Kernel PCA, which

one of the common method, uses kernel space for decomposition so that deal with non-linear properties.[2]. Multiway, multiblock, multiscale and multimode PCA also developed for increasing the model accuracy.[2–7] When PCA is constructed to sub-model, it can be more sensitive to the abnormality. Furthermore, recently knowledge based methodologies or qualitative model based methodologies are integrated with these dimension reduction model to verify the knowledge and accelerate the model accuracy. Neural network and Bayesian networks also integrated with PCA for supervised learning for knowledge based theory.[8–10]

Even though, enormous data and developed algorithms, various fault is still occurred. It is because data and the algorithms are focused on detecting the fault from operation data variation difference between normal and abnormal operation. To monitor the various conditions, multi-mode operation monitoring methods should be developed to detect fault accurately. Also most fault diagnosis analysis is based on the specific historic fault data. This method is proper for frequently fault but when the new variation fault, it does not working at all. Therefore, it is needed that extract the root cause information between normality and abnormality independent with the historic fault data,

1.2 Research objectives

The objective of this thesis is to propose a root cause information at initial fault stage for real time diagnosis. Especially, to provide an accurate root causality even when the fault has a small intensity and only using the normal operation data. To develop this objective, fault variation is defined from normal operation data to solve the new fault. Yoon et al., MacGregor el al. and Yue and Qin proposed fault direction which associate the normal intensity and fault magnitude. This vector is used for detecting fault and enhancing the fault information from historic data.

From these concept, proposed methodology comprises 2 parts. First is normal modeling and monitoring part and second is fault detection and root cause diagnosis part.

The objective of normal modeling and monitoring part is to manage the detail normal operation mode through k-means clustering, k-nearest neighbors (kNN) and minimum distance to mean (MDM) algorithms. k-means clustering method divides overall normal operation data into k using global T-score which from global principal component analysis (global PCA). Separated normal operation mode used for detail normal PCA model, which call local PCA model. New sample data can be classified proper local normal operation data from (MDM) and (kNN).[9] These procedure is described in detail in chapter 3.

The objective of fault detection and root cause diagnosis part is to provide root cause information at the same time the fault was detected. When the fault is detected,

PCA select proper subspace, which defined principal component subspace (PCS) or residual subspace(RS), so that removes normal portion in fault data.[11] And then, singular value decomposition convert the contribution to fault magnitude. Finally, multivariate granger causality method calculated causality based on time series data. From the causality, root cause in initial fault stage can be provided fault information. To verify the performance, the suggested method is applied liquefied natural gas (LNG) plant fractionation process dynamic model.[8,12] From the fault scenario in this dynamic model, accuracy of the detection performance and root cause analysis is evaluated.

1.3 Outline of the thesis

The outline of thesis is organized as follows. Chapter 1 introduce research motivation and objective. Chapter 2 describes the background theories that used in algorithms development. In chapter 3, the fault detection procedure is proposed in detail. And then application results to the LNG fractionation process are given. Chapter 4 is comprised fault detection and root cause diagnosis algorithm. As same as chapter 3, application results to LNG fractionation process are given. Chapter 5 presents the conclusions and suggestions for future work.

CHAPTER 2 : **Methodologies of fault detection and root cause analysis**

2.1 **Introduction**

In data-driven monitoring and fault diagnosis, there are enormous methodologies have been developed recently. Among them, reduction method such as principal component analysis (PCA)[13], partial least squares (PLS)[14] and fisher discriminant analysis[1] is widely used because these are basic concept used in many applications. However, classical reduction methods have a certain limit that in nonlinear, the model accuracy decreased sharply. For these reasons, non-linear reduction methods developed such as using kernel space, high-order data structure or integrated with other non-linear method.[15–17] There are a lot of methodologies and their applications, in terms of fault diagnosis has problem. Meanwhile, there are integrated methods, data driven analysis and knowledge base or model base. Dai and Gao review the integration knowledge base and data driven.[18] Hou and Wang review briefly at model-based to data-driven control.[7]

Although many developed methodologies, a methodology that can cope with all situation is impossible. In this study, instead of developing new algorithm, it is focused on detailed multi-normal modeling to specify relation between normal and fault. For multi-normal molding, clustering and class matching methodologies are used. Clustering is used for several data set that can be grouped with meaningful

relation. Also there are variety of methodologies and their application depends on the complicity of data structure.[9] If data structure is or can be low dimensioned, simple method should be more useful for integration and tuning. For this reason, this paper focused on simple classification and patter recognition method are adjusted while simplify the data structure. In this chapter, introduce k-means clustering, minimum distance to mean (MDM), k-nearest neighbors (kNN), PCA and multivariate granger causality (MVGC) as the background theories.

2.2 k-means clustering

Clustering is the analysis that binding the scattered data, which have same dimension, into meaningful group by a certain criteria. There are many developed algorithms in clustering, in this work, k-means clustering is used for classification. In cases of low dimension, such as 1, 2 or 3-dimensions which are intuitively identifiable, this clustering method is powerful than other method in terms of visual perception.

k-means clustering method is based on euclidean distance. This concept is very simple. First step is place k points into the space randomly that assumed centroids of group. Then calculate the distance between object and centroids and assign the group that has closest centroids. All objective have been assigned, centroids move to new positions. If the new position changed objective to other group, that means unstable, repeat and find new centroids. There is no objective moving anymore, clustering is

finished. Figure 2-1 shows the procedure briefly.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad \text{Eq 2-1}$$

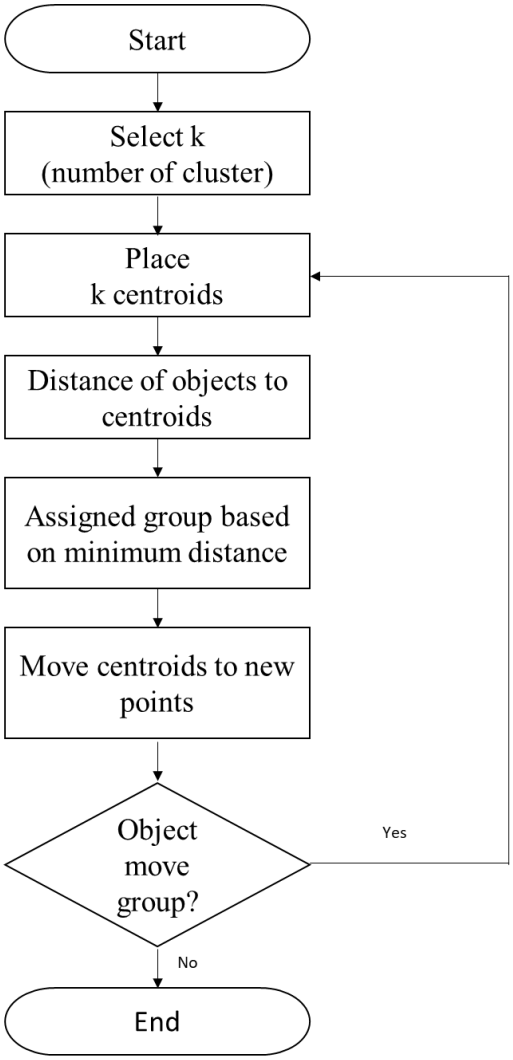


Figure 2-1 k-means clustering

2.3 Minimum distance to means and k-nearest neighbors algorithm

To match observations into proper class is essential part for modeling accuracy. In this work, operation mode, used as a class, has below 3-dimension space, as well as the reason for using k-means clustering, minimum distance to means (MDM) and k-nearest neighbors (kNN) method are used for class matching.

MDM and kNN method are based on euclidean distance. A new observation is compared the distance with normal training variables. MDM uses mean values of each class training data. The new observation has class the nearest index. In kNN method, k variables are used in the nearest order. The most class of k nearest variables is assigned to the new sample. Therefore, in order not to recognize the wrong class, it is used a sufficiently large training data and k value. Both methods are used together to reduce misreading.

2.4 Principal component analysis

PCA is a one of the classic method for reducing the dimension of a data set.[13,19] It is very powerful for large data set which have linear correlation. This method convert high dimension correlated sensors into low dimension linearly uncorrelated variables. PCA can be derived singular value decomposition (SVD) of data *X*,

$$X = U\Sigma W^T \quad \text{Eq 2-2}$$

where Σ is a $n \times m$ rectangular diagonal matrix of positive number σ_k, U is an $n \times n$ singular values of *X* matrix and *W* is an $m \times m$ matrix. In terms of factoriazation,

$$X^T X = W\Sigma^T U^T U\Sigma W^T \quad \text{Eq 2-3}$$

$$= W\Sigma^2 W^T \quad \text{Eq 2-4}$$

where Σ is the square diagonal matrix with the singular valued of *X*. From the SVD, score matrix can be given as,

$$T = XW \quad \text{Eq 2-5}$$

$$= U\Sigma W^T W \quad \text{Eq 2-6}$$

$$= U\Sigma \quad \text{Eq 2-7}$$

The process alarm limit can be defined by Hotelling's T^2,

$$T_a{}^2 = \frac{(n^2-1)}{n(n-a)} F_\alpha(a, n-a) \quad \text{Eq 2-8}$$

where *a* represents the number of selected principal components (PCs), $F_{\alpha}(a, n-a)$ is the *F*-distribution with *a* and (*n*-*a*) degrees of freedom and α means the level of significance.

In residual part, which is defined squared prediction error (SPE), can detect the fault using the Q-statistics limit (${\delta_{\alpha}}^2$), given as

$$ {\delta_{\alpha}}^2 = \theta_1 \left\{ \frac{h_0 c_{\alpha} \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{{\theta_1}^2} \right\}^{1/h_0} \quad \text{Eq 2-9} $$

$$ \theta_i = \sum_{j=a+1}^{m} {\lambda_j}^i \quad \text{Eq 2-10} $$

$$ h_0 = 1 - \frac{2\theta_1 \theta_3}{3{\theta_2}^2} \quad \text{Eq 2-11} $$

Hotelling's T^2 limit and Q-statistics limit are the process threshold of normal operation. This two methods are complementary to detect fault but should used separately. When the process is out of the Hotelling's T^2limit, the fault space is defined as principal component subspace (PCS). Likewise when the SPE alarm goes off, the process fault is defined as residual subspace (RS).

New sample vector, which means the real time data for monitoring and diagnosis, can be projected two parts, principal component subspace (PCS) and residual subspace(RS).[20]

$$ x = \hat{x} + \tilde{x} \quad \text{Eq 2-12} $$

$$ \hat{x} = PP^T x \quad \text{Eq 2-13} $$

$$\tilde{x} = (I - PP^T)x \qquad \text{Eq 2-14}$$

$\hat{x}$ is PCS projection and $\tilde{x}$ is RS projection. $\hat{x}$ and $\tilde{x}$ have own monitoring variable in each subspace.

2.5 Multivariate Grange causality

Granger causality (GC) is based on linear autoregressive modelling of stochastic process.[17,21,22] Briefly, if a variable $X_2(t)$ has an information of a future $X_1(t)$ variable and there is no information that other series used in the predictor, then $X_2(t)$ is said a 'granger cause' $X_1(t)$. This concept is interpreted as shown,

$$X_1(t) = \sum_j^p A_{11}(j)X_1(t-j) + \sum_j^p A_{12}(j)X_2(t-j) + \mathcal{E}_1(t) \qquad \text{Eq 2-15}$$

$$var(\mathcal{E}_1(t)) = \Sigma_1 \qquad \text{Eq 2-16}$$

$$X_2(t) = \sum_j^p A_{21}(j)X_1(t-j) + \sum_j^p A_{22}(j)X_2(t-j) + \mathcal{E}_2(t) \qquad \text{Eq 2-17}$$

$$var(\mathcal{E}_1(t)) = \Sigma_2 \qquad \text{Eq 2-18}$$

where *A(j)* is AR coefficient, k is model order, $\mathcal{E}(t)$ is prediction errors. These equations are the definition of full regression bivariate AR model. If there is no dependence between $X_1(t)$ and $X_2(t)$, $A_{12}(j)$ and $A_{21}(j)$ are 0. This concept is consideration of the reduced regression.

$$X_1(t) = \sum_j^p A'_{11}(j)X_1(t-j) + \mathcal{E}_{1(2)}(t) \qquad \text{Eq 2-19}$$

$$var\left(\mathcal{E}_{1(2)}(t)\right) = \Sigma_{1(2)} \quad \text{Eq 2-20}$$

$$X_2(t) = \sum_{j}^{p} A'_{22}(j)X_2(t-j) + \mathcal{E}_{2(1)}(t) \quad \text{Eq 2-21}$$

$$var\left(\mathcal{E}_{2(1)}(t)\right) = \Sigma_{2(1)} \quad \text{Eq 2-22}$$

$\Sigma_{i(j)}$ means variance $\mathcal{E}$ at restricted j. GC from $X_2(t)$ to $X_1(t)$ is defined as log-likelihood ratio,

$$\mathrm{F}_{X_2 \to X_1} = ln\frac{\Sigma_{1(2)}}{\Sigma_1} \quad \text{Eq 2-23}$$

This concept can be extended to m variables system by AR-coefficients given as eq 2-19.[21]

$$\begin{Bmatrix} X_1(t) \\ X_2(t) \\ \vdots \\ X_m(t) \end{Bmatrix} = \sum_{k=1}^{\infty} \begin{bmatrix} A_{11}(k) & A_{12}(k) & \cdots & A_{1m}(k) \\ A_{21}(k) & A_{22}(k) & \cdots & A_{2m}(k) \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1}(k) & A_{m2}(k) & \cdots & A_{mm}(k) \end{bmatrix} \begin{Bmatrix} X_1(t-k) \\ X_2(t-k) \\ \vdots \\ X_m(t-k) \end{Bmatrix} + \begin{Bmatrix} \varepsilon_1(t) \\ \varepsilon_2(t) \\ \vdots \\ \varepsilon_m(t) \end{Bmatrix}$$

Eq 2-24

GC from $X_j(t)$ to $X_i(t)$ is given eq.20

$$\mathrm{F}_{X_j \to X_i} = ln\frac{\Sigma_{i(j)}}{\Sigma_i} \quad \text{Eq 2-25}$$

In this multivariate form, Σ_i means cov(ε_i),which is defined variance from all other *m* variables, and $\Sigma_{i(j)}$ represent cov($\varepsilon_{i(j)}$), which is defiend *(m-1)* variables that restricted *j*.

In this work, 'MVGC tool box', developed by Barnett et al. in matlab code, is used for analyze.[23]

CHAPTER 3 : Multi-mode monitoring using k-means clustering, minimum distance to mean, k-nearest neighbors and principal component analysis

3.1 Introduction

Process monitoring and fault diagnosis have been a significantly important part of plant management recently. Accumulated data can be enhancement of monitoring and fault diagnosis performance. However, the amount of the data is so much existing, it is impossible to use all of these data fully or effectively.[24] Therefore it is important that how to select and use your data.

Principal component analysis (PCA), which widely used in multivariate statistical process control (MSPC), is used as the concept of how to use. Most of plant use the PCA for monitoring fault detection. And also it is used for diagnosis information that providing the affected sensors. Lane et al. adjusts to film manufacturing process for monitoring and information of affected sensors[25], Li et al. use recursive PCA to thermal annealing process for monitoring.[26] Garcia et al., Gallagher et al., used multi-way PCA for batch process so that the quality management from best case.[27,28]

It is important to use the methods integrated others so that increasing the accuracy. There are another ways to increase the accuracy of model. One of them is overall model dived sub-model for detail monitoring and diagnosis. This method can be modeling the segment of process variables or time variables. In terms of process

variables, MacGregor et al. proposed multiblock PCA and partial least square (PLS) and Westerhuis et al., Smilde et al., use this method.[19,29,30] They divide variables from using their criteria, so increase the accuracy of model performance. In terms of time variables, Lu et al, Zhao et al., Zhu et al are considered time variables at modeling stage. They separate the operation mode or time stage so that manage the process rigorously.

Liquefaction natural gas (LNG) fractionation process, which used in study for validation, has many operation mode because they have a feature of the downstream process and affected by refrigeration and liquefaction process. Therefore, it is important for monitoring to separate the operation mode. For the separate the normal operation mode, k-means clustering method is adjust to T-score, derived from global PCA, used for classification. For the new sample data matching with training data, minimum distance to mean (MDM) and k-nearest neighbors (KNN) method are used. From the time segment data, local data, is modeled by PCA so that process monitoring. From the local PCA monitoring, compared with global PCA monitoring and univariate monitoring, have a good performance.

This chapter is comprised 5 section. Section 3.1 is introduction. Section 3.2 is described k-means clustering, MDM and kNN. In section 3.3, LNG plant fractionation process dynamic model and fault scenario described. In section 3.4, the result of 45 scenario described and section 3.5 is conclusion.

3.2 Multimode-PCA monitoring integrated with k-means clustering, minimum distance to mean and k-nearest neighbors

The proposed monitoring method is consist of 2 parts. First is normal operation data modeling. Because PCA is based on linearity system, it is necessary to adjust a linear interpretation of the process. Among the ways to attempt linearity, in these study, PCA modeling as segmented normal process data was applied. Therefore, statistical method and clustering method are integrated with PCA in modeling procedure. Second is process multi-mode monitoring. PCA projection, minimum distance to mean (MDM), k-nearest neighbors (kNN) and contribution charts are used for process monitoring.

3.2.1 Normal operation data modeling

Handling normal data is of great importance to modeling process. The performance of a model depends on how data is selected and preprocessed. In this study, Normal operation data modeling has 3 parts.

First step is global PCA modeling. PCA method is shown in section 2.4. Any normal data contains noise and disturbance. For removing them, outlier data is eliminated by 3-sigma rule. This rule removes data in excess of 99.7% of the normal data range. It is important for improving model accuracy. And then these data are rearranged to the scaled data using from mean and standard deviation. After the preprocessing,

these dataset decompose 2 or 1 principal component by the PCA method. It is defined as global modeling that construct model using from overall normal data. Figure 3-2 shows these procedure.

Second step is k-means clustering to divide global data into several local data. Accumulated data, which have same dimension, have high probability of similar dynamic behavior. If data with similar dynamic behavior divided and classification own them, these information can increase the performance accuracy of model. k-means clustering is described in section 2.2.

From global modeling, T-score variable are derived. Generally, T-scores are 1~3 dimensional chart, so process state or behavior can be understood intuitively. For example, figure 3-9 shows the global normal data state in 1-D chart. Visually and intuitively, this chart shows 3 normal states in them. These scattered data can be classified to several clusters by k-means clustering. From this clustering method, normal data are divided into k-class normal data.

Final step for normal modeling is local PCA modeling. Here, except for using local class normal data, separated data go through same process. Outlier of local normal data is removed by 3-sigma rule. And then they are rearranged by scaling. Lastly, PCA decompose them to reduced variables. Each local PCA model calculated own process limit, Hotelling's T^2 and Q-statistic.

Figure 3-1 shows overall procedure of normal operation data modeling. Figure 3-2,3,4 shows global PCA modeling procedure, operation mode classification and local

PCA modeling procedure respectively.

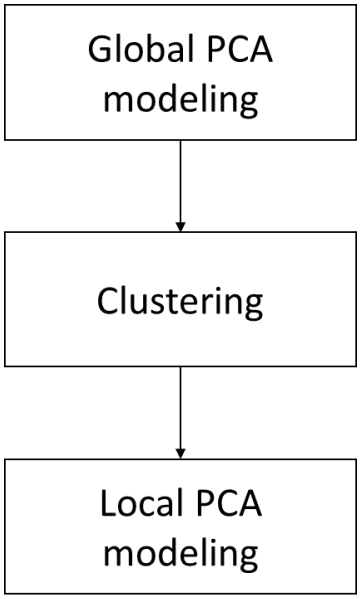


Figure 3-1 Overall normal operation modeling procedure

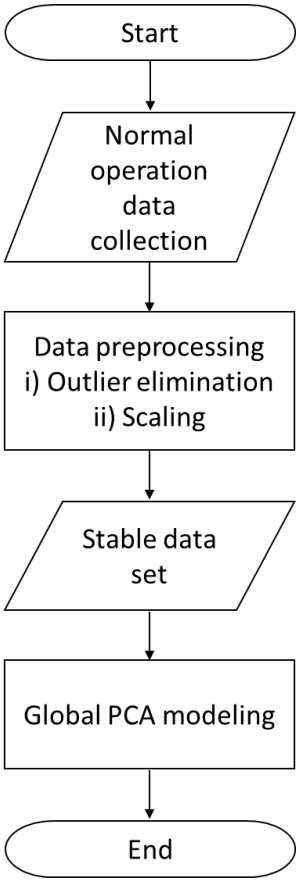


Figure 3-2 Global PCA modeling procedure

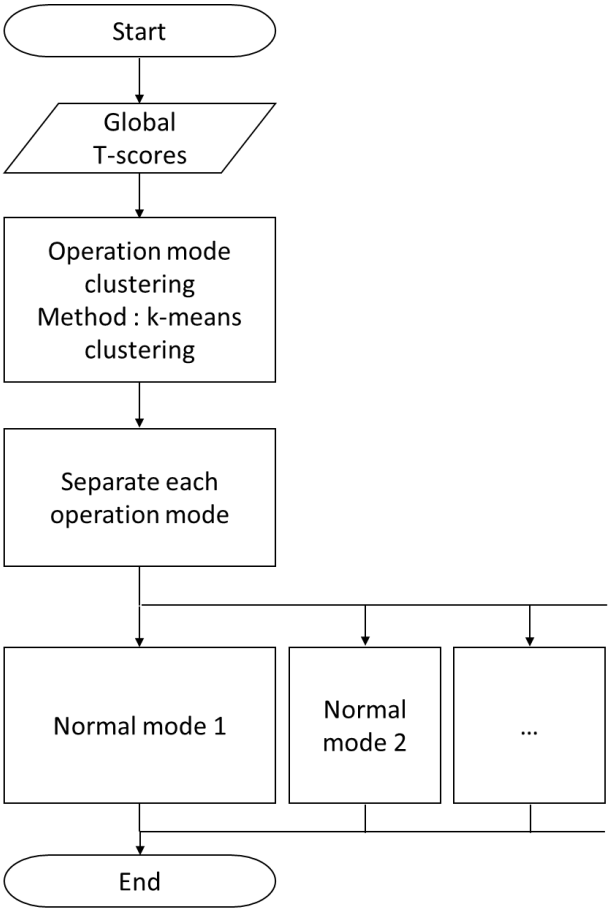


Figure 3-3 Operation mode classification procedure

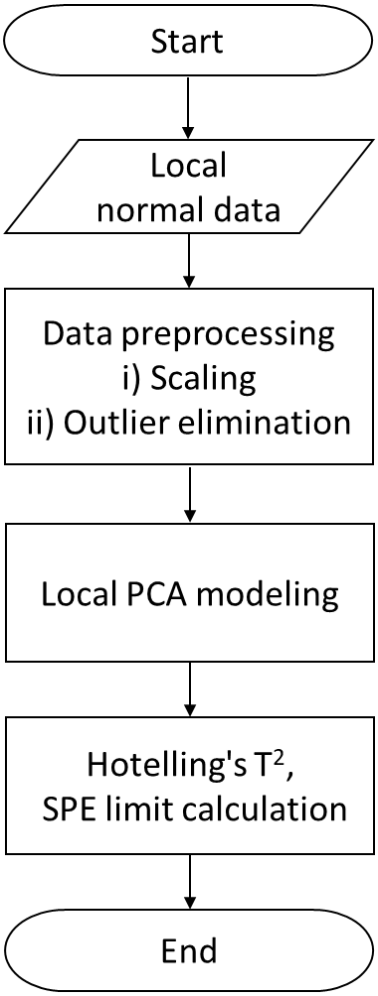


Figure 3-4 Local PCA modeling procedure

3.2.2 **Process multi-mode monitoring**

The proposed normal modeling is used for multi-mode monitoring. This part has 2 steps. First step is matching of the new sample data with the class classified k-means clustering. In this study, minimum distance to means (MDM) and k-nearest neighbors (kNN) are used to match the class. MDM and kNN method are described in section 2.3.

For classification, it is needed to processing new sample data. New sample data, which means real time data, must rearranged by global normal scale data. After that, this data is projected to global PCA space. From the space, T-score of new sample is derived. Using these T-score, MDM and kNN match the new sample with local normal mode that the most similar to new sample. This procedure is shown in figure 3-5.

After the class is decided, the new sample data go through the scaling using from own local normal class scaling factors. Scaled data can be projected into local PCA space where they are decomposed to Hotelling's T^2 and Q-statistic. If there is nothing occurred, this new sample is defined as a normal data so that is is saved in own local data class. However the new sample data are occurred the alarm, contribution is analyzed for identifying affected sensors. If alarm is Hotelling's T^2, T-contribution should be used and SPE contribution should be used if Q-statistics alarm goes off. This procedure is shown in figure 3-6.

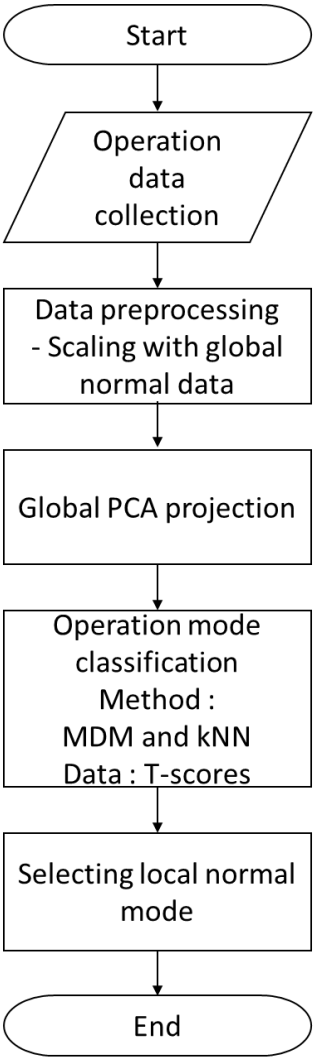


Figure 3-5 Operation class matching procedure

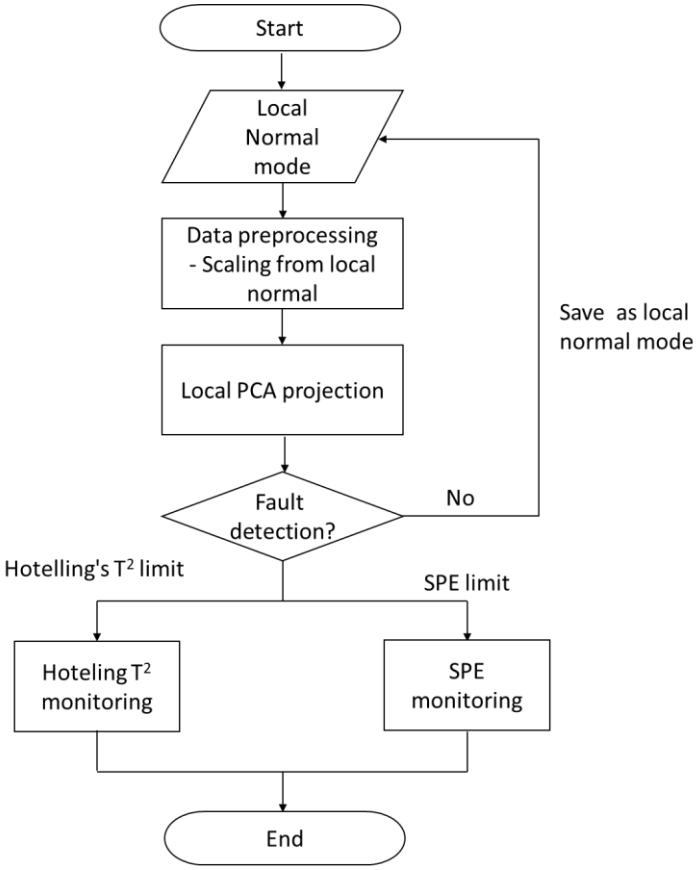


Figure 3-6 Local monitoring procedure

3.3 Liquefied Natural Gas (LNG) fractionation process

3.3.1 Model description

Liquefied Natural Gas (LNG) fractionation process is one of the major process in LNG plant, 3 others are pre-treatment process, liquefaction process and storage (shipment). This process separates mixed refrigerant for purification. It is consist of 4-main column, deMethanizer, deEthanizer, dePropanizera and deButanizer. Each column separates methane, ethane, propane and butane respectively. These columns have sensitive low temperature and high pressure because of small carbon material properties. Therefore, for verifying the accuracy of proposed algorithm, this process is developed dynamic model using from Aspen hysys® v8.1 simulator. Figure 3-7 represent schematic of LNG fractionation process. Table 3-1, 2 show the monitoring variables and descriptions.

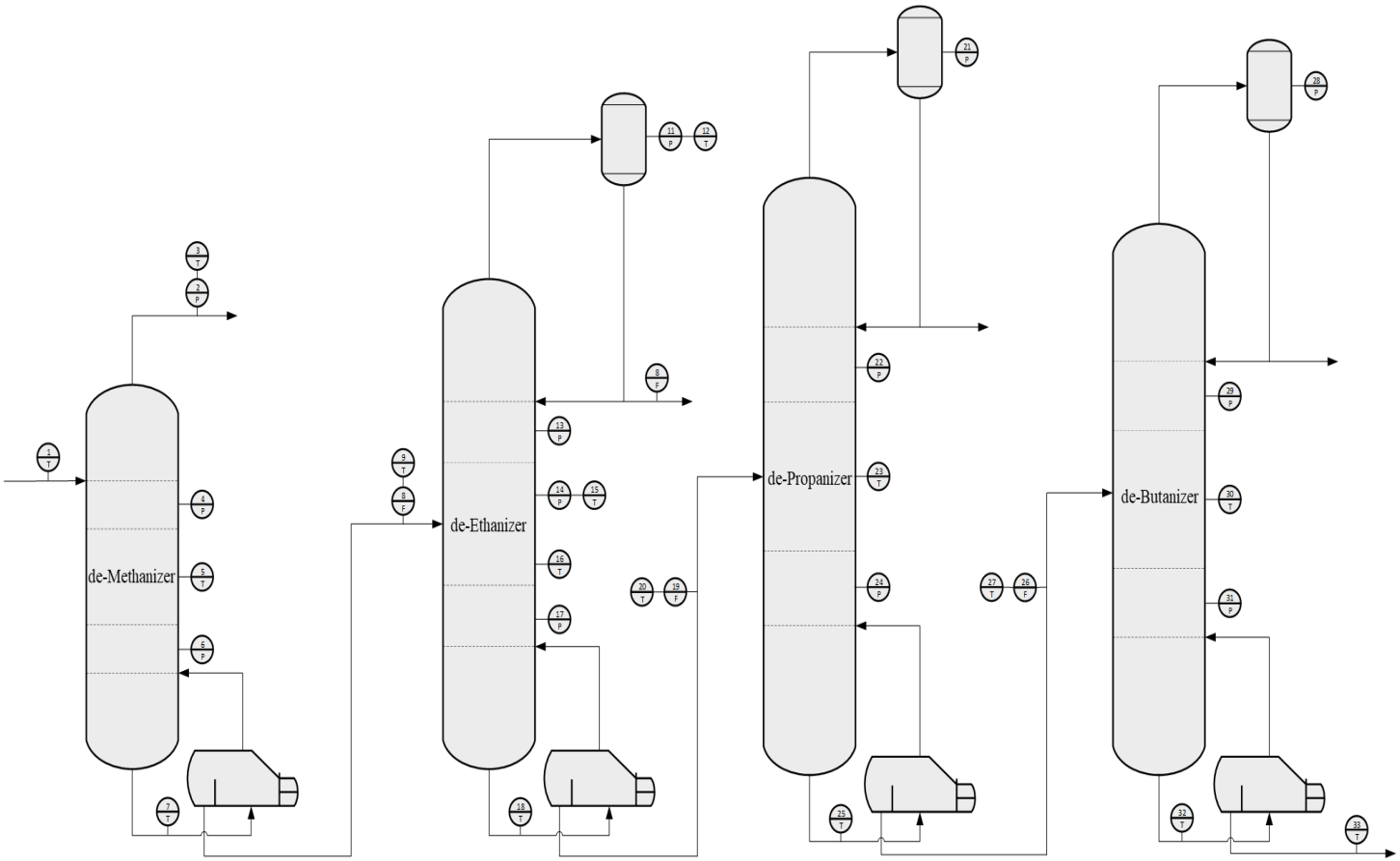


Figure 3-7 schematic of LNG fractionation process

Table 3-1 Monitoring variables and descriptions

Stream No.	Tag name	Tag description
1	dM-in-T	deMethanizer inlet stream temperature
2	dM-top-P	deMethanizer top stream pressure
3	dM-top-T	deMethanizer top stream temperature
4	dM-1st-P	deMethanizer stage-1 pressure
5	dM-7st-T	deMethanizer stage-7 temperature
6	dM-13st-P	deMethanizer stage-13 pressure
7	dM-reb-T	deMethanizer reboiler temperature
8	dE-in-F (dM-bot-F)	deEthanizer inlet stream flow rate (deMethanizer bottom stream flow rate)
9	dE-in-T (dM-bot-T)	deEthanizer inlet stream temperature (deMethanizer bottom stream temperature)
10	dE-top-T	deEthanizer top stream temperature
11	dE-cond-P	deEthanizer condenser pressure
12	dE-cond-T	deEthanizer condenser temperature
13	dE-1st-P	deEthanizer stage-1 pressure
14	dE-11st-P	deEthanizer stage-11 pressure
15	dE-11st-T	deEthanizer stage-11 temperature
16	dE-20st-T	deEthanizer stage-20 temperature
17	dE-28st-P	deEthanizer stage-28 pressure

Table 3-2 Monitoring variables and descriptions

Stream No.	Tag name	Tag description
18	dE-reb-T	deEthanizer reboiler temperature
19	dP-in-F (dE-bot-F)	dePropanizer inlet flow rate (deEthanizer bottom stream flow rate)
20	dP-in-T (dE-bot-T)	dePropanizer input temperature (deEthanizer bottom stream temperature)
21	dP-cond-P	dePropanizer condenser pressure
22	dP-1st-P	dePropanizer stage-1 pressure
23	dP-19st-T	dePropanizer stage-19 temperature
24	dP-37st-P	dePropanizer stage-37 pressure
25	dP-reb-T	dePropanizer reboiler temperature
26	dB-in-F (dP-bot-F)	deButanizer inlet flow rate (deButanizer bottom stream flow rate)
27	dB-in-T (dP-bot-T)	deButanizer input temperature (deButanizer bottom stream temperature)
28	dB-cond-P	deButanizer condenser pressure
29	dB-1st-P	deButanizer stage-1 pressure
30	dB-17st-T	deButanizer stage-17 temperature
31	dB-34st-P	deButanizer stage-34 pressure
32	dB-reb-T	deButanizer reboiler temperature
33	dB-bot-T	deButanizer bottom stream temperature

3.3.2 **Normal and fault scenario description**

The condition LNG fractionation process varies according to the previous process, such as MR-process or liquefaction. Therefore, this process has several narrow condition mode. In this work, 3 normal modes are applied; (A) has initial stream temperature -16.3℃ ~ -15.3℃ and pressure 61.0bar ~ 62.0bar, (B) has initial stream temperature -16.8℃ ~ -15.8℃ and pressure 61.5bar ~ 62.5bar, (C) has initial stream temperature -17.8℃ ~ -16.8℃ and pressure 62.5bar ~ 63.5bar. These normal mode are simulated in stable convergence area and suitable for product specification area. Table 3.3 and 3.4 show the process overall specifications and normal modes condition.

Table 3-3 Process overall specification

	C1[%]	C2[%]	C3[%]	C4[%]	others
deMethanizer top stream	91.60	5.31	2.06	0.80	0.23
deEthanizer top stream	0.00	99.55	0.45		
dePropanizer top stream	0.00	0.00	99.61	0.39	
deButanizer top stream	0.00	0.00	1.23	98.00	0.77

Table 3-4 3 normal modes condition

Normal case	Process initial stream Temperature [℃]	Process initial stream Pressure[bar]
A	-16.3 ~ -15.3	61.0 ~ 62.0
B	-16.8 ~ -15.8	61.5 ~ 62.5
C	-17.8 ~ -16.8	62.5 ~ 63.5

In column, there are coexistence of gas phase, liquid phase, vaporization and liquefaction. For that reason, column processes have many problem with temperature.[31] In this work, 3 types of temperature abnormal situation and 2 leaking fault are supposed. Fault (1) is deEthanizer inlet (deMethanizer bottom) flow leaking, (2) is dePropanizer inlet (deEthanizer bottom) flow leaking, (3) is deEthanizer reboiler temperature overheating, (4) is deEthanizer condenser temperature overcooling, (5) is deMethanizer reboiler temperature overheating. Each fault has 3 different strengths; 3%, 5%, 10% intensity compared with normal condition. Overall, there are 3 normal mode, 5 types of fault and 3 types of strength that total 45 cases are generated and used for analysis. All fault case consist of 500 seconds normal data and 2000 seconds abnormal data. Table 3-5 shows fault scenarios. Figure 3-8 shows fault location in schematic process diagram.

Table 3-5 Fault scenarios

Fault type	Fault intensity	Fault description
1	3%	deEthanizer inlet(deMethanizer bottom) flow 3% leaking
	5%	deEthanizer inlet(deMethanizer bottom) flow 5% leaking
	10%	deEthanizer inlet(deMethanizer bottom) flow 10% leaking
2	3%	dePropanizer inlet(deEthanizer bottom) flow 3% leaking
	5%	dePropanizer inlet(deEthanizer bottom) flow 5% leaking
	10%	dePropanizer inlet(deEthanizer bottom) flow 10% leaking
3	3%	deEthanizer reboiler temperature 3% overheating
	5%	deEthanizer reboiler temperature 5% overheating
	10%	deEthanizer reboiler temperature 10% overheating
4	3%	deEthanizer condenser temperature 3% overcooling
	5%	deEthanizer condenser temperature 5% overcooling
	10%	deEthanizer condenser temperature 10% overcooling
5	3%	deMethanizer reboiler temperature 3% overheating
	5%	deMethanizer reboiler temperature 5% overheating
	10%	deMethanizer reboiler temperature 10% overheating

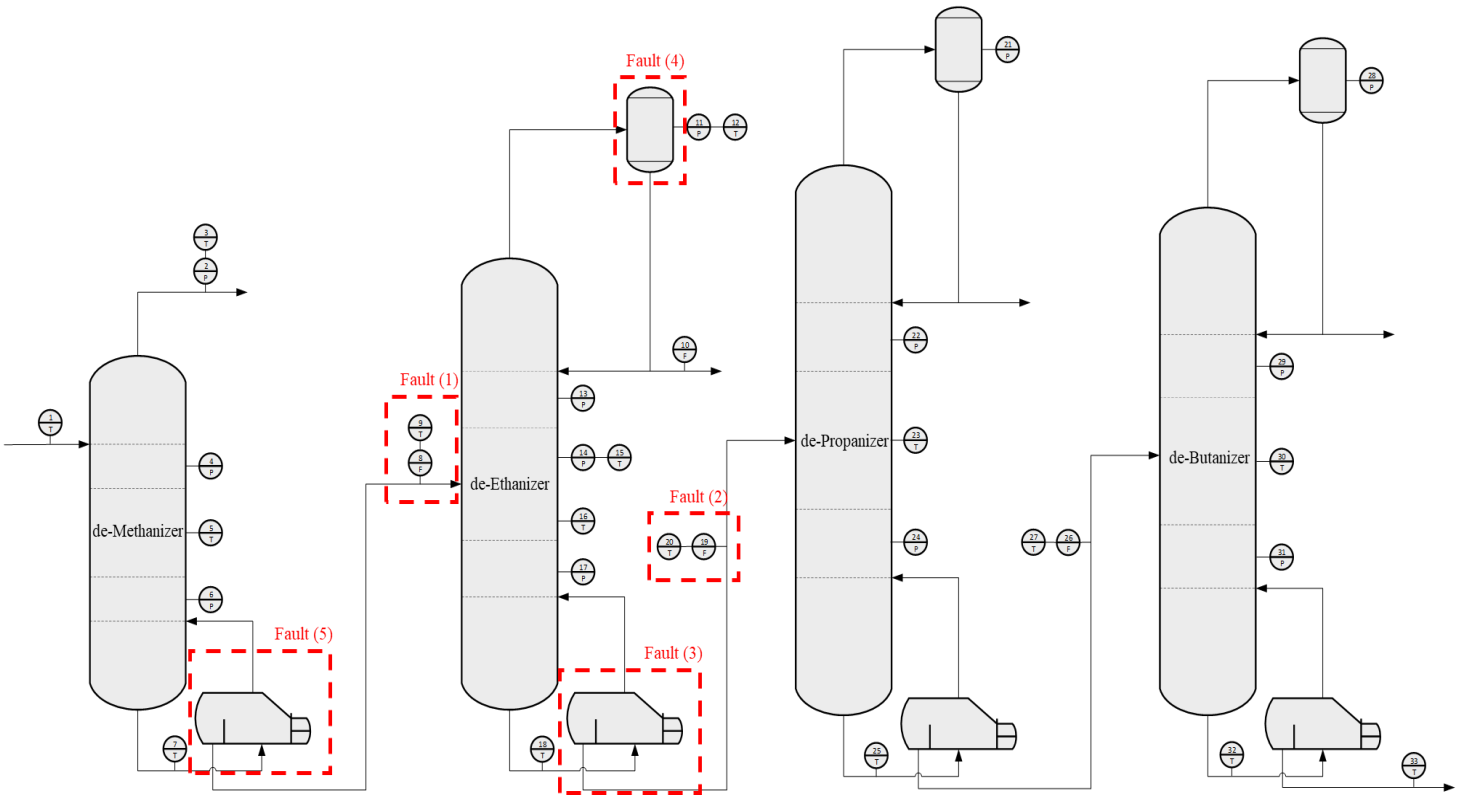


Figure 3-8 Fault location in schematic process diagram.

3.4 **Results**

3.4.1 **Multi-mode modeling**

First of all, overall normal data go through the normal modeling procedure, which is described in section 3.2. In this step, overall 120,000 seconds are used for clustering. There are 3 types of normal operation data which have 40,000 seconds data respectively. From the figure 3-9, 3 types of normal operation data can be recognized. With a factor k of 3, k-means clustering classified the global normal into 3 types of local normal for multi-mode modeling. 3 local normal data are treated outlier elimination and scaling same as global normal data. After preprocessing, PCA decomposes each local data to reduced space. Finally, process limit for monitoring that Hotelling's T^2 and SPE.

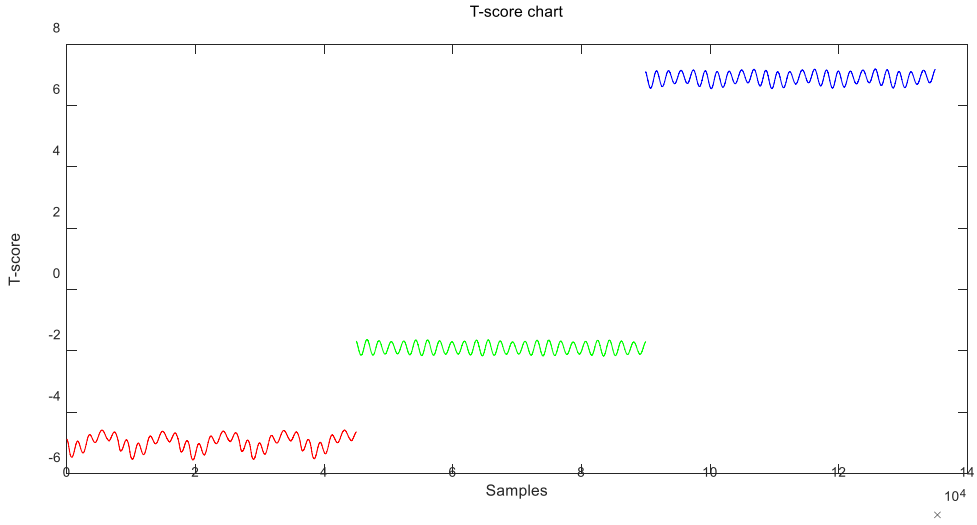


Figure 3-9 k-means clustering result

3.4.2 **Monitoring fault detection**

For performance analysis of the proposed multi-mode monitoring method, this method is compared with shewart 3-sigma method and global PCA. Fault detection accuracy (FDA) and fault detection rate (FDR), which are broadly used in monitoring performance, are adjusted to 3 monitoring methods. FDA and FDR are defined form type I and type II errors as shows in figure 3-10 and equation 3-1,2.

$$FDA = \frac{TP + TN}{TP + FN + FP + TN} \quad \text{Eq 3-1}$$

$$\text{FDR} = \frac{TN}{FP + TN} \quad \text{Eq 3-2}$$

		Real class	
		Normal	Fault
Predicted class	Normal	True Positive(TP)	False Positive(FP)
	Fault	False Negative(FN)	True Negative(TN)

Figure 3-10 Type I & II errors

Results are organized in table 3-6 ~ 3-9. The proposed monitoring method are improved better performance than global PCA in all cases. 43 cases out of 45 showed better performance than global PCA. Only two cases, B-1-10% and C-1-10% are the same results in FDA and FDR. Because multi-mode PCA has a specific limit line, it can detect the fault more sensitively than global PCA.

Compared with the univariate 3-sigma method, 43 cases out of 45 are detected same or faster in multi-mode PCA. Especially, 3% intensity faults are detected remarkably faster than shewart 3-sigma. This is because small fault changes the relevant variables, it may not be able to exceed the individual variable limits. Therefore, univariate monitoring method does not detect until the fault grows. However, because the local PCA integrates the variation in individual variables, it can detect the fault faster and more accurately.

Table 3-6 FDA and FDR results in Shewart 3-sigma chart, Global PCA and Multi-mode PCA-Normal A part

		Shewart 3-sigma		Global PCA		Multi-mode PCA	
Fault type	Intensity	FDA [%]	FDR [%]	FDA [%]	FDR [%]	FDA [%]	FDR [%]
1	3%	65.08	56.35	64.48	55.60	87.40	84.25
	5%	87.40	84.25	74.40	68.00	94.60	93.25
	10%	99.96	99.95	99.92	99.90	99.96	99.95
2	3%	94.64	93.30	91.80	89.75	95.00	93.75
	5%	98.00	97.50	96.36	95.45	97.84	97.30
	10%	99.68	99.60	99.36	99.20	99.64	99.55
3	3%	88.64	85.80	80.32	75.40	90.44	88.05
	5%	95.84	94.80	88.24	85.30	96.08	95.10
	10%	97.96	97.45	93.00	91.25	97.96	97.45
4	3%	74.32	67.90	-	-	93.44	91.80
	5%	89.56	86.95	-	-	96.24	95.30
	10%	96.80	96.00	-	-	98.28	97.85
5	3%	87.28	84.10	92.08	90.10	93.60	92.00
	5%	93.60	92.00	95.76	94.70	96.80	96.00
	10%	97.80	97.25	98.08	97.60	98.60	98.25

Table 3-7 FDA and FDR results in Shewart 3-sigma chart, Global PCA and Multi-mode PCA-Normal B part

		Shewart 3-sigma		Global PCA		Multi-mode PCA	
Fault type	Intensity	FDA [%]	FDR [%]	FDA [%]	FDR [%]	FDA [%]	FDR [%]
1	3%	86.24	82.80	58.40	48.00	90.88	88.60
	5%	89.76	87.20	69.48	61.85	99.80	99.75
	10%	99.96	99.95	99.96	99.95	99.96	99.95
2	3%	99.64	99.55	87.64	84.55	99.64	99.55
	5%	99.72	99.65	94.60	93.25	99.72	99.65
	10%	99.80	99.75	98.64	98.30	99.80	99.75
3	3%	96.48	95.60	77.52	71.90	99.16	98.95
	5%	98.40	98.00	86.84	83.55	99.56	99.45
	10%	99.16	98.95	91.04	88.80	99.64	99.55
4	3%	96.32	95.40	-	-	98.44	98.05
	5%	97.76	97.20	-	-	98.96	98.70
	10%	99.08	98.85	-	-	99.56	99.45
5	3%	86.36	82.95	92.88	91.10	95.24	94.05
	5%	92.60	90.75	96.20	95.25	97.40	96.75
	10%	97.80	97.25	98.28	97.85	98.80	98.50

Table 3-8 FDA and FDR results in Shewart 3-sigma chart, Global PCA and Multi-mode PCA-Normal C part

		Shewart 3-sigma		Global PCA		Multi-mode PCA	
Fault type	Intensity	FDA [%]	FDR [%]	FDA [%]	FDR [%]	FDA [%]	FDR [%]
1	3%	37.04	21.30	-	-	82.60	78.25
	5%	99.96	99.95	79.88	74.85	99.96	99.95
	10%	99.96	99.95	99.96	99.95	99.96	99.95
2	3%	90.88	88.60	80.84	76.05	93.08	91.35
	5%	95.24	94.05	85.68	82.10	96.24	95.30
	10%	96.44	95.55	88.28	85.35	97.48	96.85
3	3%	96.32	95.40	75.04	68.80	98.92	98.65
	5%	98.56	98.20	86.04	82.55	99.52	99.40
	10%	99.24	99.05	90.64	88.30	99.64	99.55
4	3%	95.80	94.75	-	-	97.52	96.90
	5%	97.40	96.75	-	-	98.44	98.05
	10%	98.84	98.55	88.68	85.85	99.24	99.05
5	3%	87.60	84.50	89.76	87.20	94.08	92.60
	5%	93.80	92.25	94.48	93.10	96.80	96.00
	10%	98.00	97.50	97.60	97.00	98.60	98.25

3.5 Conclusion

This study proposes a monitoring method for early detection using k-means clustering, kNN, MDM and PCA. First, in normal modeling procedure, T-scores calculated from global PCA are classified into k-normal operation using k-means clustering. Next these local normal operations are modeled as multi-mode PCA for detail monitoring fault detection. The new samples are assigned class by MDM and kNN. From assigned local PCA projection, FDR and FDA evaluate the result compared with global PCA and shewart 3-sigma method. From the result, proposed method has better performance in all cases than global PCA. Only 2 out of 45 cases are the same result and the others com out with increased performance. When compared to shewart 3-sigma method,

In monitoring part, MDM and kNN methods are used for matching the proper local normal so that system is monitored by local normal PCA modeling. From the classification, proposed method detected the fault faster than global PCA and shewart 3-sigma. In 2 cases are shown slow FDR and inaccuracy FDA, but there is little difference. 8 cases are the same result and 35 cases have good performance in FDR and FDA. The average FDA of proposed method is improved by about 5~10% to 97%, where global PCA is 88% and shewart 3-sigma is 93%. The average FDR of proposed method is also improved by about 5~10% to 96%, where global PCA is 85% and shewart is 91%.

CHAPTER 4 : Root cause analysis at early abnormal stage using principal component analysis and multivariate Ganger causality

4.1 Introduction

Nowadays, the development of physical sensing technology, distributed control system (DCS) and computing technology have brought about the development of plant scale. As a result, these huge processes make out an enormous amount of data. They enable detailed analysis about system for maximize production and minimize safety costs. However, various information and accumulated data are not always guaranteed the efficiency and the safety. There are reasons that the sensors have complex relationship between each variable, malfunction, calibration error, missing, etc. If such incorrect information is provided under abnormal situation, it causes confusion in the analysis of causes and problem-solving. Under the assumption that there are no physical error, the key of process management is an intuitive interpretation of numerous sensors and correlations between them. When it comes to fault occurred, early detection and analysis of root cause are the major interest area for efficiency and safety.

Multivariable statistical process control is the conventional data-based methodology for monitoring and fault detection. It defines the normal state, including steady state, that all process variable are operated in approximately the same position as normal state. In industrial area, univariate monitoring method is commonly used in plant. There are a lot of sensors managed and monitored by

operator. They each have several limits for their own purpose that control limit, warning limit, risk limit, etc. However, these univariate methods are needed skilled operators, knowledge of process and time for analysis. For this reason, multivariate monitoring method is developed in academic area and have being adjusted to real plant. Principal component analysis (PCA) is one of the most preferred method in various system, chemical plant, steel industry, fuel cell, batch process etc. This method decompose a large number of sensors to a small number of component as maintaining the origin information. It uses orthogonal projection for converting of correlated variables into linearly uncorrelated variables. The normal state in process are defined that these data are used for PCA modeling. There are two monitoring variables whether they are fault or not, Hotelling's T^2 and Q-statistic. Hotelling's T^2 indicates the distance between center which reduced dimension of normal state and observation which projected onto reduced space. SPE indicates dimension-reducing distance between PCA normal model and observation. Therefor these two indicators are monitored simultaneous. When the fault occurred, each contribution data give information about the affected variables in fault state. These data can be useful in root cause diagnosis.[16,32] Accordingly, the two indicators have great strengths in a quick detection, visualization and diagnosis information.

The PCA methodology is applied to various industries. DOFASCO, which is known as steel industry, casting and desulphurization process are adjusted PCA for early detection and visualization.[33–35] This company uses PCA in real time online monitoring system. They use Hotelling's T^2 and SPE plot for monitoring and these contribution chart for diagnosis. Although it has strength in early detection and diagnosis generally, it also has malfunction or fails in contribution. This is happened

when the normal PCA modeling includes so many sensors which are noised and unimportant that screening the state of process.

In Jeong et al. resolved this problem integrating the factor analysis and PCA.[36] This method sort out these disturbing sensors in Molten Carbon Fuel Cell in order to get accurate fault detected time and diagnostic contribution. For the detail normal PCA modeling, multi-mode PCA method are developed. It is integrated hierarchical clustering and PCA for global PCA. Jiang et al. proposed Bayesian interference and joint probability integrated with PCA that adjust training and identify the various sub-block normal modes.[37] Ha et al. used k-nearest neighbors for matching the local normal mode and adjusted PCA for detection.

Most in case, PCA is developed and integrated with other methodologies for monitoring efficiency like early detection. However, it is as important for fast monitoring as root cause analysis. In generally, analysis of root cause depends on historic data, qualitative knowledge or expert system.[38–42] Although historic data is very enormous amount in data storage system, they are mainly normal data or different process condition compared with present condition. Knowledge base qualitative analysis or expert system are very accurate on the one hand. However they have a major weakness. Abnormal situation is various depending on the condition, therefore there are too many cases to analyze advance. Also, it take a lot of time to analyze after the fault. Resolving these problems, data driven fault detection and root cause diagnosis are developed recently. MacGregor and Kourti, Yue et al. and Qin proposed a reconstruction which integrated Hotelling's T^2 and Q-statistic approach for increasing accuracy of fault direction and diagnosis root cause.[20,43] Recently methods of diagnosis a root cause with this concept is studied. Ahnmed

et al. used singular value decomposition (SVD) for amplifying root cause variables.[44,45] Using the residual contribution fault direction, which modeled by historical fault data, enhance the contribution data for propagation path. Kitano et al. also used reconstruction-contribution from historic fault data.[46] Enhancing contribution from fault direction has good performance at high intensity and frequently occurring abnormalities. However, in a small fault, these methods can not shows good performance. Also, because it depends on the historical fault data, it is difficult to give a root cause information when a new faults occurs.

In this work, it is focused on root cause analysis about new fault and initial fault stage. PCA method detects the abnormal state using Hotelling's T^2 and Q-statistic. When the fault is detected, its contribution data are scaled and analyzed by SVD so that it should be find the sensors affected by fault. These sensors are used in multivariate granger causality (MVGC) method. Granger causality (GC) is widely used for root causality between sensors. This method based on vector autoregressive model(VAR), which is linearly regress model.[23,45] This method just need time series data at specific situation when it is identified fault or abnormal. Especially, it indicates better efficiency in using the key variables. As mentioned above, integrated MVGC and the selected variables from the PCA and SVD make the effective performance in root cause diagnosis.

This paper is divided into four major section. The first section describes theory about PCA and contribution handling. In section 2, methodologies are proposed; modeling, fault detection and root cause diagnosis. The next section describes LNG fractionation process for case study and fault scenario briefly. And then, result and discuss about fault scenario. Finally, the last section presents conclusion.

4.2 Monitoring and root cause diagnosis

In this study, PCA, SVD and MVGC are used for root cause diagnosis. These methodologies are generally based on linearity system. To attempt linearity, abnormal data is reconstructed at specific time interval, where enough short compared to the time interval in normal modeling.

4.2.1 Fault magnitude sensors

Background of PCA is same as shown in Section 2.4.

When a fault is occurred, a new sample data is divided into normal and abnormal. These two portion is also reflected in PCS and RS respectively. It is expressed

$$x = x^* + \Xi_i f \quad \text{Eq 4-1}$$

$$\hat{x} = \hat{x}^* + \hat{\Xi}_i f \quad \text{Eq 4-2}$$

$$\tilde{x} = \tilde{x}^* + \tilde{\Xi}_i f \quad \text{Eq 4-3}$$

where Ξ_i represents the fault direction, $\hat{\Xi}_i$ and $\tilde{\Xi}_i$ are the fault directions on PCS and RS, respectively, and i refers to the number of principal components. The strength of the fault is represented by $\|f\|$, which changes over time. Generally, the portion of contribution about $\hat{x}^*$ and $\tilde{x}^*$ is insignificant compared with the fault strength. Therefore, the contribution of x is about the same as the contribution of $\Xi_i f$. However, it is difficult to ignore that very small intensity and the initial stage of fault.[45] Therefore, the normal variation embedded in the fault data need to be removed or minimized. For this purpose, the statistics of the normal contribution data, which is used to train data in PCA modeling, should be used to scale the fault

contribution data. They are given as

$$\hat{x}' = \hat{\Xi}_i f \quad \text{Eq 4-4}$$

$$\tilde{x}' = \tilde{\Xi}_i f \quad \text{Eq 4-5}$$

Whether $\hat{\Xi}_i$or $\tilde{\Xi}_i$is determined by which alarm is triggered. These two variables only works in their respective subspace. These two parameters are reconstructed using from singular value decomposition (SVD). When the alarm occurred, fault data set, that has an $k \times m$ in which k is samples corresponding to m sensors, is expressed.

$$X_i = [\, x_1 \;\; x_2 \;\; \dots \;\; x_k]^T \quad \text{Eq 4-6}$$

It can be interpreted from the eq 4-14 or eq 4-15, given as

$$\hat{X'}_i^{\,T} = \hat{\Xi}_i[\, f_1 \;\; f_2 \;\; \dots \;\; f_k] \quad \text{Eq 4-7}$$

$$\tilde{X'}_i^{\,T} = \tilde{\Xi}_i[\, f_1 \;\; f_2 \;\; \dots \;\; f_k] \quad \text{Eq 4-8}$$

These equation represent PCS and RS respectively. $\hat{X'}_i^{\,T}$or $\tilde{X'}_i^{\,T}$is convert to the covariance matrix to analyze the covariation among contributions,

$$\mathrm{Cov}\left(\hat{X'}_i^{\,T}\right) = \left[\hat{\sigma}_{pq}\right] \;\; ; \;\; p, q = 1,2, \dots, k \quad \text{Eq 4-9}$$

$$\mathrm{Cov}\left(\tilde{X'}_i^{\,T}\right) = \left[\tilde{\sigma}_{pq}\right] \;\; ; \;\; p, q = 1,2, \dots, k \quad \text{Eq 4-10}$$

SVD method adjust to covariance matrix $\hat{X'}_i^{\,T}$or $\tilde{X'}_i^{\,T}$so that convert correlate variables into uncorrelate vairbles while retaining the singular values. SVD decomposes the covariance matix into an orthgonal matrix (U_i), diagonal matrix (D_i) and transpose of orthogonal matrix V_i^T. Matrix U_i has 3 important features. First,

U_i consists of fault point data. Second, it is removed or minimized the normal portion. Finally, it is decomposition values of process. For this reason, the values in first column of U_i are meaning the contributions that make up the process state. $\hat{\Xi}_i$ and $\tilde{\Xi}_i$ are given as eq 4-13, eq 4-14.

$$\hat{X'}_i^T = \hat{U}_i \hat{D}_i \hat{V}_i^T \quad \text{Eq 4-11}$$

$$\tilde{X'}_i^T = \tilde{U}_i \tilde{D}_i \tilde{V}_i^T \quad \text{Eq 4-12}$$

$$\hat{\Xi}_i = \hat{U}_i(:,1) \quad \text{Eq 4-13}$$

$$\tilde{\Xi}_i = \tilde{U}_i(:,1) \quad \text{Eq 4-14}$$

Finally, fault magnitude sensors, that hierarchical sensors, are selected by the procedure shown in Figure 4-2. From the absolute fault magnitude data, 32% is selected as hierarchical sensors for multivariate Granger causality analysis, representing those that have contributions larger than the sum of the mean and the 1-sigma value.

$$\bar{u} = average(abs(U_i(:,1))) \quad \text{Eq 4-15}$$

$$\bar{\sigma} = stdv(abs(U_i(:,1))) \quad \text{Eq 4-16}$$

$$abs(U_i(:,1)) > \bar{u} + \bar{\sigma} \quad \text{Eq 4-17}$$

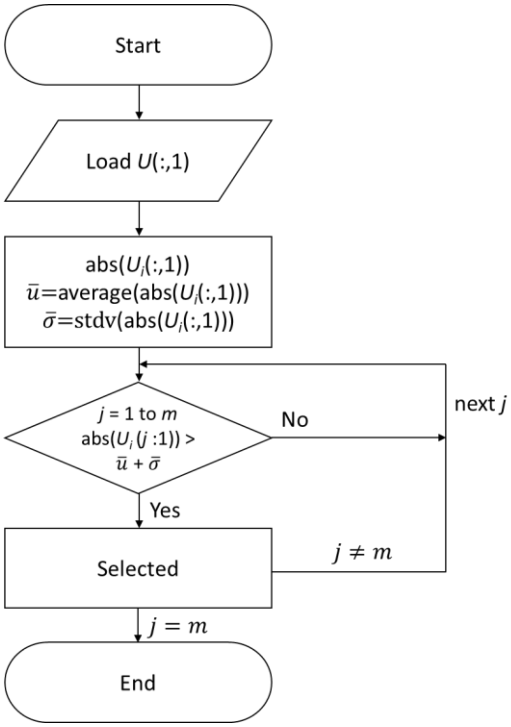


Figure 4-1 Selecting procedure of hierarchy sensors

4.2.2 **Normal modeling**

The methodologies described above are integrated for early detection and root cause diagnosis. This algorithm is divided 2 parts. First step is normal modeling so that PCA method constructs the process limit and information of normal contribution in PCS and RS respectively. Normal operation data, which already these data are known as normal state, are gathered. These data go through preprocessing. Outlier data is eliminated by 3-sigma method in statistics, which means out of 99.7% normal data ranges removed for model accuracy. And they are scaled by average and standard deviation. After stable data set is ready, PCA method decompose them to reduced spaces, PCS and RS. In these 2 spaces respectively, it is calculated that limits for monitoring and scale statistics of contribution. Figure 4-2 shows procedure of handling the normal data.

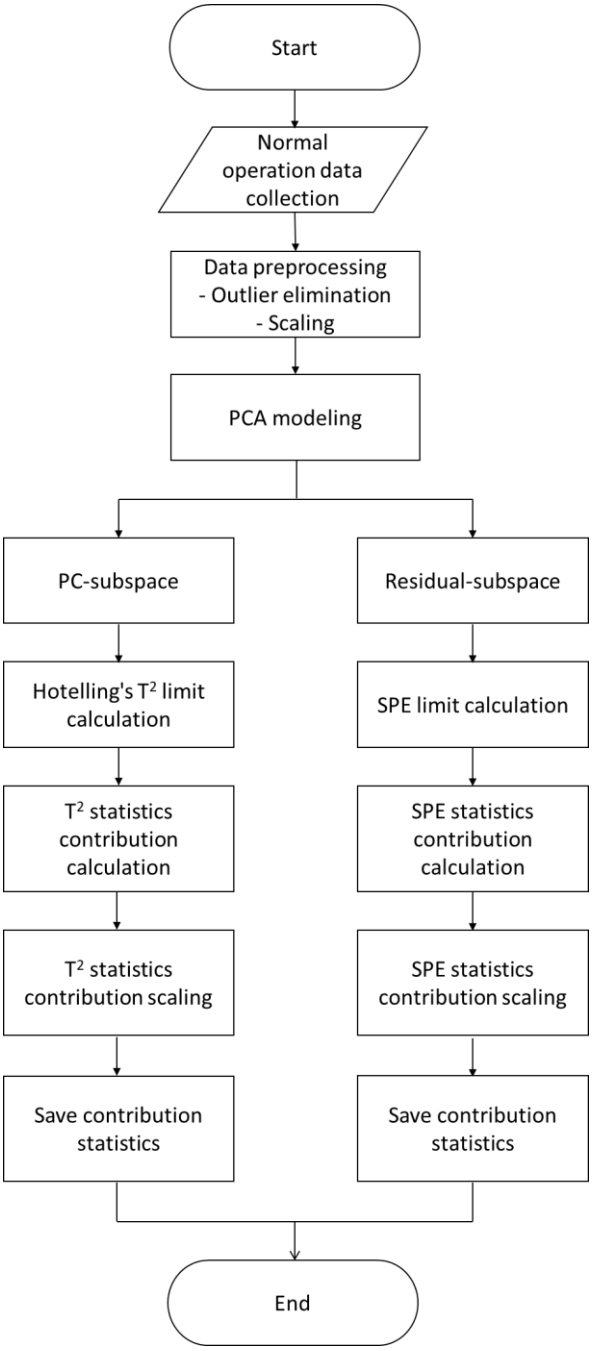


Figure 4-2 Normal data monitoring and handling procedure

4.2.3 **Fault detection and diagnosis**

Fault detection and diagnosis consist of two parts. The first part is monitoring the process. New data, real-time data, goes through the scaling process used in normal data statistics. Then, the scaled data are projected to the PCA normal model. From the projection, the model is monitored by the Hotelling's T^2 and SPE. When the process alarm occurs, the fault data are collected. Then, the subspace is determined by which alarm occurred. If the sample data exceed the Hotelling's T^2 at *k* time, the data are gathered from $(t-n)$ time to *t* time, n times before *t* time to *t* time is reached. These data goes into PC-subspace. In this space, the T-contributions are calculated and scaled by normal contribution, and then, the covariance of these contribution data is analyzed by SVD method to select the hierarchical sensors. The fault magnitude method removes the normal portion, $U_i(:,1)$, with the empirical rule selecting sensors with more than 32%. Finally, the MVGC method is performed for these sensors, resulting in the construction of a causality matrix. If the alarm occurs from SPE, the process is the same as PCS, except that it is performed in RS. This algorithm is described in Figure 4-3.

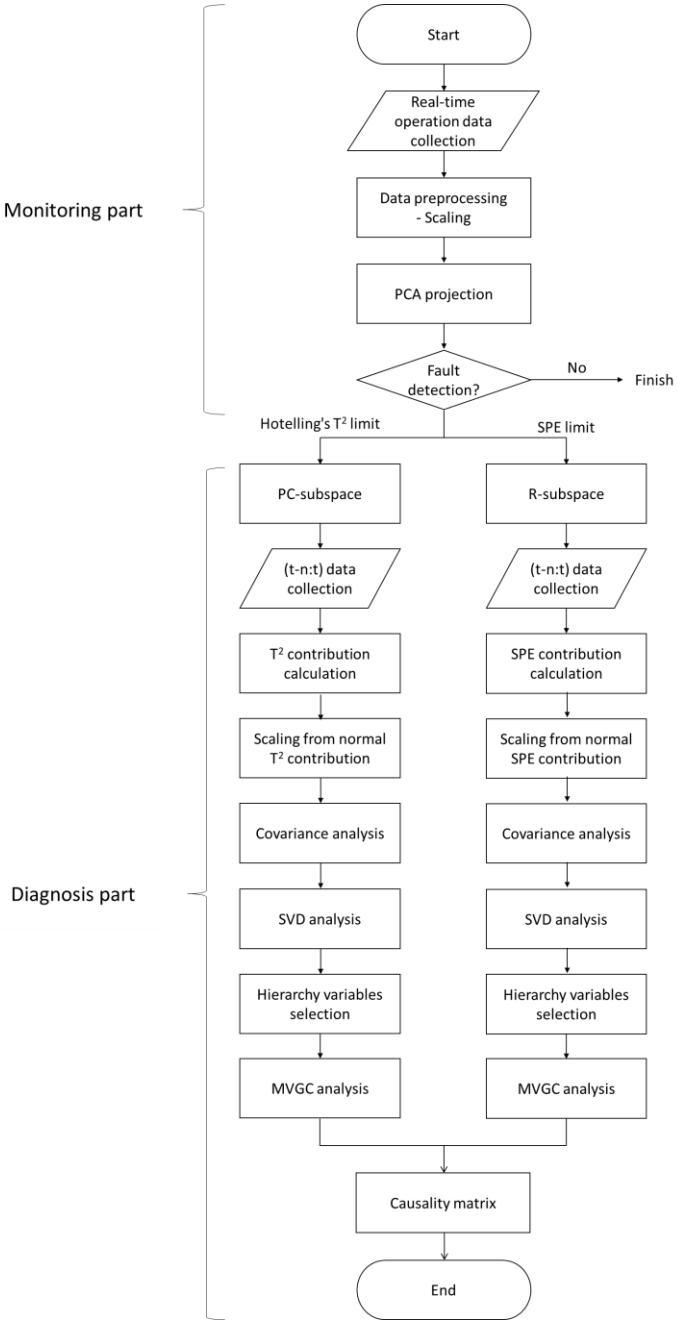


Figure 4-3 Fault detection and root cause diagnosis procedure

4.3 Application to the Liquefied Natural Gas (LNG) fractionation Process

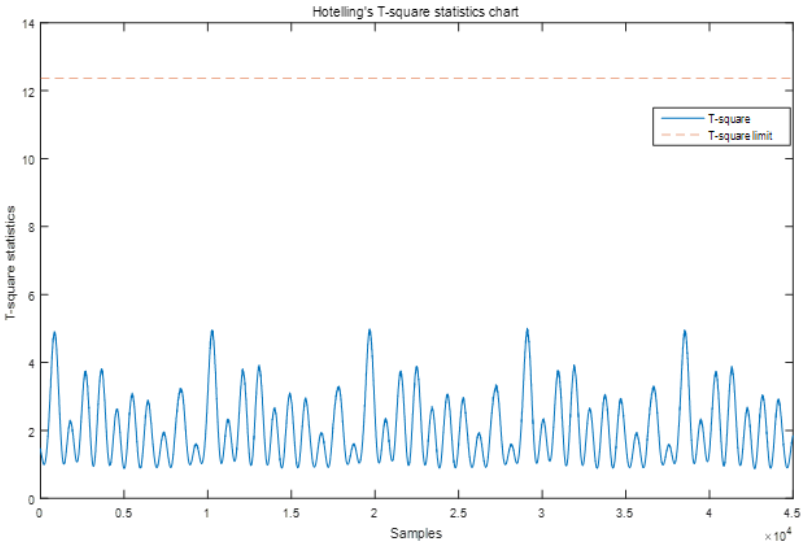
4.3.1 Process Description

LNG fractionation process has 4 main columns. (Figure 3-8) Each column separates methane, ethane, propane and butane respectively. These columns are operated in very low temperature and high pressure. Therefore, product specifications are very sensitive. Safety management is also very important. In this work, in order to generate data for verifying the algorithm, dynamic model is developed using from Aspen hysys® simulator. From this model, normal operation data is generated. Process condition, sensors are the same as shown in section 3.3.

4.3.2 Normal data processing

PCA modeling calculate in order of procedure figure 4-4. From the normal data modeling, it derive PCS and RS and its normal contribution normalization result respectively. Figure 4-4(a) and Figure 4-4 (b) shows normal data variation in PCS, Hotelling's T^2 chart and normal data variation in RS, SPE.

(a)



(b)

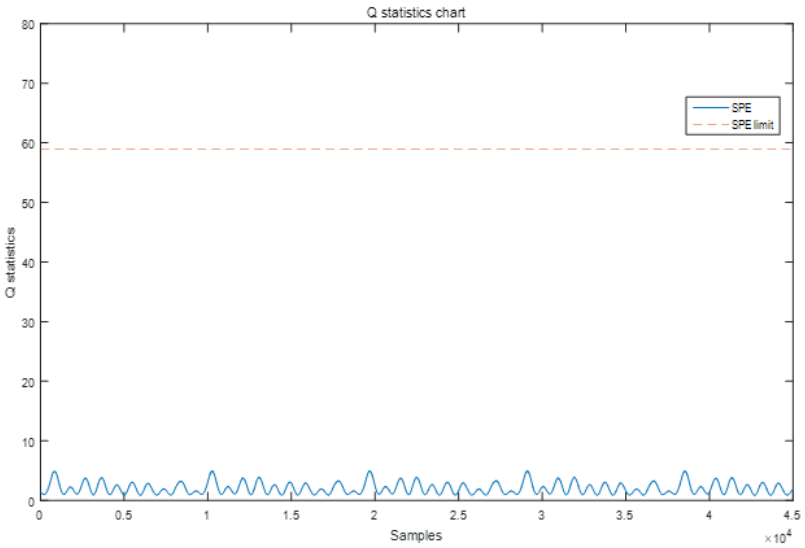


Figure 4-4 (a) Hotelling's T^2 chart (b) SPE chart at normal state

4.3.3 **Fault scenario 1 : A-3-3%**

First scenario is reboiler overheating in deEthanizer column. It is common fault in deEthanizer. This fault is caused several reasons such as control limit changed improperly, not detected in accurate temperature, malfunction in control logic, steam valve opening or human error, etc.[31,47] To simulate this fault, heat duty of deEthanizer reboiler is increased 3 % than normal state. 3% of heat duty is very small compared with the normal value, but if it lasts, there can be flooding occurred in column. Figure 4-5 (a) and (b) shows the alarm in Hotelling's T^2 and Q-statistics. Comparing figures 4-5 (a) and (b), detecting time in PCS is 240.0 seconds and RS is 289.0 seconds. PCS space detects alarm earlier than RS, analysis should be in PCS. Before using the developed algorithm, contribution chart is used for diagnosis information at PCA methodology. This is generally performed in conventional PCA analysis.

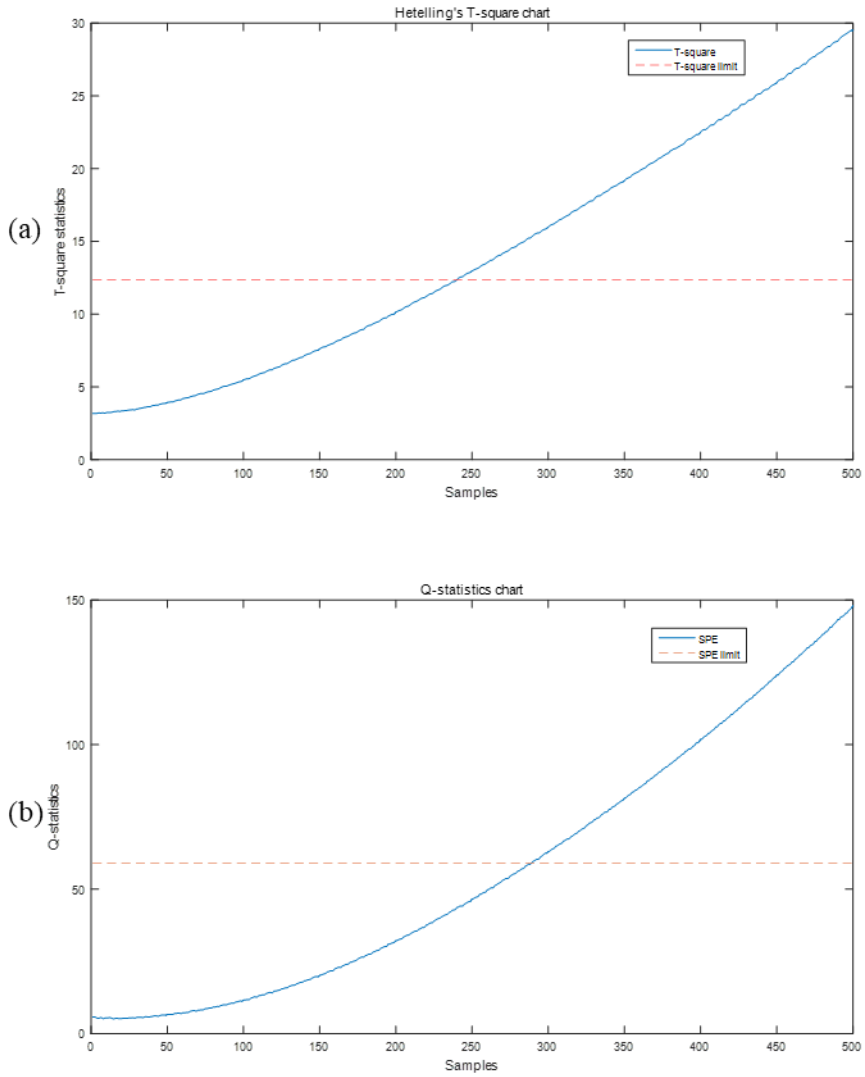


Figure 4-5 (a) Hotelling's T^2 chart (b) SPE chart at 0~500 seconds fault occurred

From the figure 4-5, alarm detected earlier in Hotelling's T^2 than Q-statistics, therefore T-contribution in PCS should be used for diagnosis. To demonstrate the excellence of the algorithm, conventional PCA contribution and RS amplification are compared with the developed algorithm. Generally, a contribution chart is used at alarm time for finding sensors affected by fault. Figure 4-6 shows the T-contribution chart at alarm time. If the contribution chart provide proper information about reboiler overheating, temperature sensors around the deEthanizer reboiler, such as dE-reb-T, dP-in-T (dE-bot-T), or dE-20st-T should provide larger values than those from the other sensors. However, it is pointing to irrelevant variables, such as dM-in-T, dM-top-T and dP-reb-T, which are higher than group of deEthanizer sensors on the whole. This contribution trends is due to intensity of fault that means small scale fault or initial stage, therefore they take up more portion of normal information than fault information.

The RS amplification method uses the fault direction, which is constructed from the historical SPE contribution of same fault, to enhance affected sensors. [44,45] Using the RS amplification method, its result is shown as Figure 4-7. This method causes dE-reb-T and dP-in-T (dE-bot-T) to be sufficiently higher than the other sensors. However dM-in-T, dM-top-T, dB-in-F (dP-bot-F), which the next large sensors, are screening the relevant sensors that are affected by fault. This is because RS information is used even though alarm is occurred in PCS. These screening sensors acts as misleading elements in MVGC analysis. It can be shown in root analysis

result, table 4-2 and figure 4-11.

Compared with conventional PCA contribution and RS amplification, the developed algorithm shows a much more reasonable result, as indicated in Figure 4-8. The sensors of deEthanizer mainly affected by fault are increased, such as dE-20st-T, dE-reb-T, and dP-in-T (dE-bot-T). In addition, portions of unrelated fault sensors are decreased, such as dM-in-T, dM-top-T, and dP-cond-P. This is because ignoring a portion of the normal contribution enhances the related sensors and diminishes the normal behavior of the unaffected sensors, such that the hierarchical sensors can stand out.

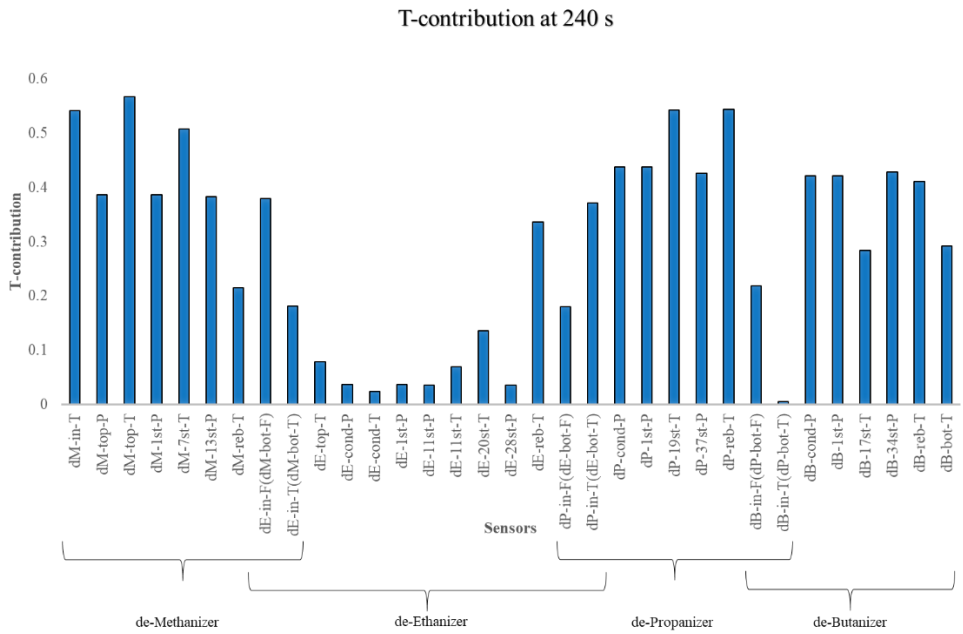


Figure 4-6 T-contribution chart at alarm occurred

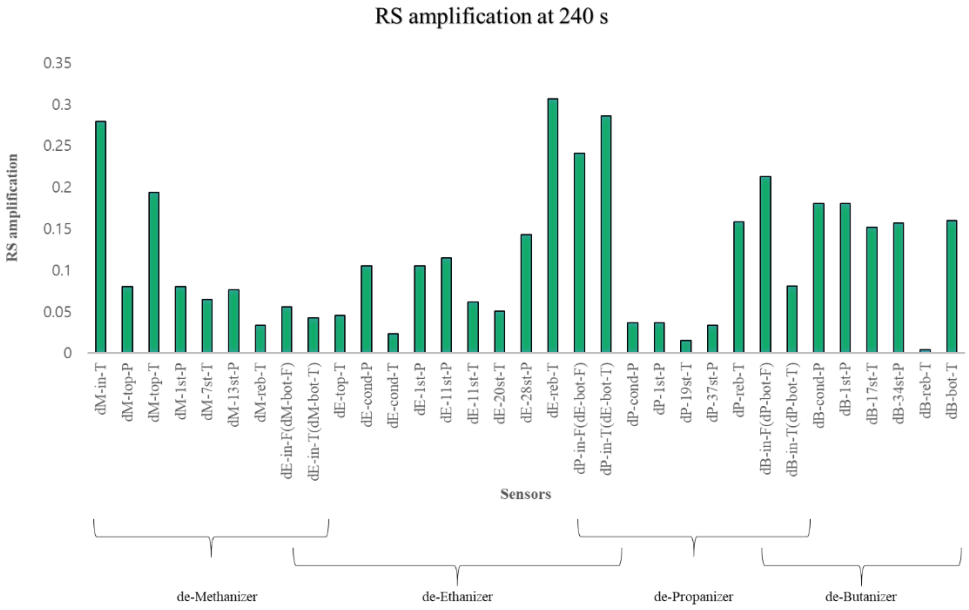


Figure 4-7 RS amplification chart at alarm occurred

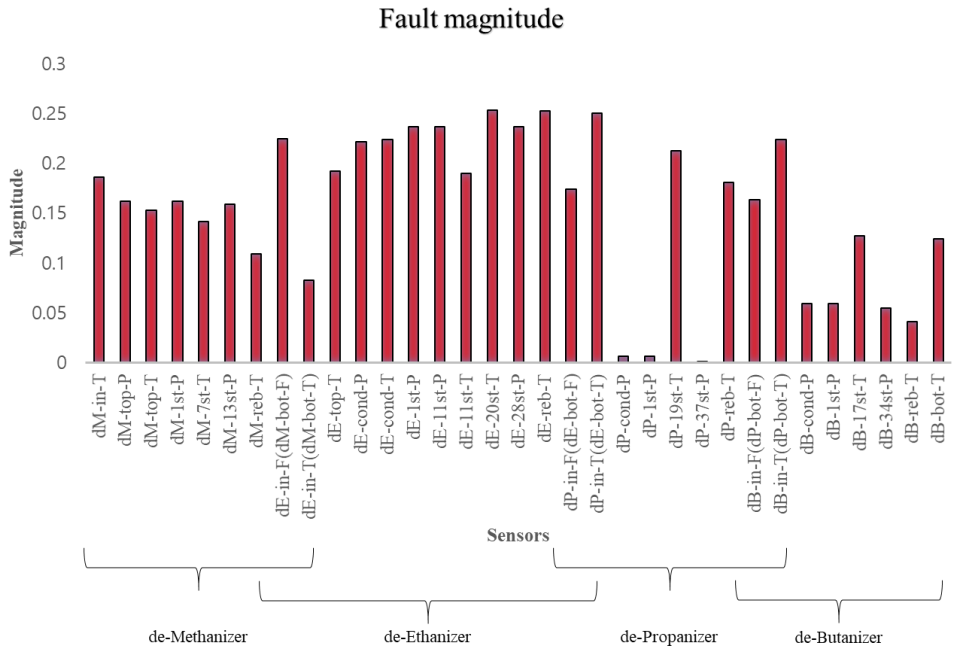


Figure 4-8 Fault magnitude chart at 141 ~ 240 seconds

Using the method mentioned in Figure 4-2, the magnitude sensors are selected for analysis in the MVGC method. From these sensors, MVGC analysis constructs the causality matrix. This matrix shows that proposed method can provide the root causality more clearly than the RS amplification method can. The causality matrix is described in Table 4-1 and Table 4-2 for the RS amplification method and fault magnitude method, respectively. Row variables means causal sensors and column variables are the affected sensors. A significant amount of data, through a comparison of internal data in its own table, is shaded thickly, and slightly larger data shaded thinly.

Table 4-1 Granger causality using RS amplification method

	dE-reb-T	**dP-in-T (dE-bot-T)**	**dM-in-T**	**dP-in-F (dE-bot-F)**	**dB-in-F (dP-bot-F)**	**dM-top-T**
dE-reb-T	-	0.14272	0.097082	0.064009	0.015276	0.000247
dP-in-T (dE-bot-T)	0.051614	-	0.010908	8.77E-05	0.017597	5.62E-06
dM-in-T	0.056997	0.015298	-	0.001025	0.024647	0.000534
dP-in-F (dE-bot-F)	0.037005	0.012381	0.00244	-	0.008263	0.008511
dB-in-F (dP-bot-F)	0.037484	0.02653	1.32E-04	0.008476	-	6.69E-05
dM-top-T	0.004426	0.448781	0.422897	0.098121	0.010157	-

Table 4-2 Granger causality using proposed method

	dE-20st-T	**dE-reb-T**	**dP-in-T (dE-bot-T)**	**dE-1st-P**	**dE-11st-P**	**dE-28st-P**
dE-20st-T	-	0.26911	0.03337	0.03353	0.00799	0.00943
dE-reb-T	0.00011	-	0.21870	0.01235	0.01291	0.00086
dP-in-T (dE-bot-T)	0.00281	0.00585	-	0.02228	0.02003	0.04079
dE-1st-P	0.00007	0.00000	0.00052	-	0.00006	0.00338
dE-11st-P	0.00000	0.00000	0.00029	0.00406	-	0.01191
dE-28st-P	0.00020	0.00035	0.00385	0.00153	0.00403	-

The proposed method shows the greatest strength of cause and effect from dE-reb-T to dE-20st-T. Next is from dP-in-T (dE-bot-T) to dE-reb-T. They are reasonable processes from the interpretation point of view because the initial root causes are located near the reboiler. On the other hand, RS amplification shows the main causality is from dP-in-T (dE-bot-T) and dM-in-T to dM-top-T. Intuitively, these causalities are not reasonable because the physical distance between the sensors is too far or these are not sensor relationships that can affect or be affected. These compared root causalities are described in Figure 4-9. For a visual comparison, the fault magnitude method and RS amplification method are expressed in the process diagram, as shown in Figure 4-10. Green arrows indicate the RS amplification causality, and red arrows represent the fault magnitude causality. The developed method should be recognized more clearly in the process diagram.

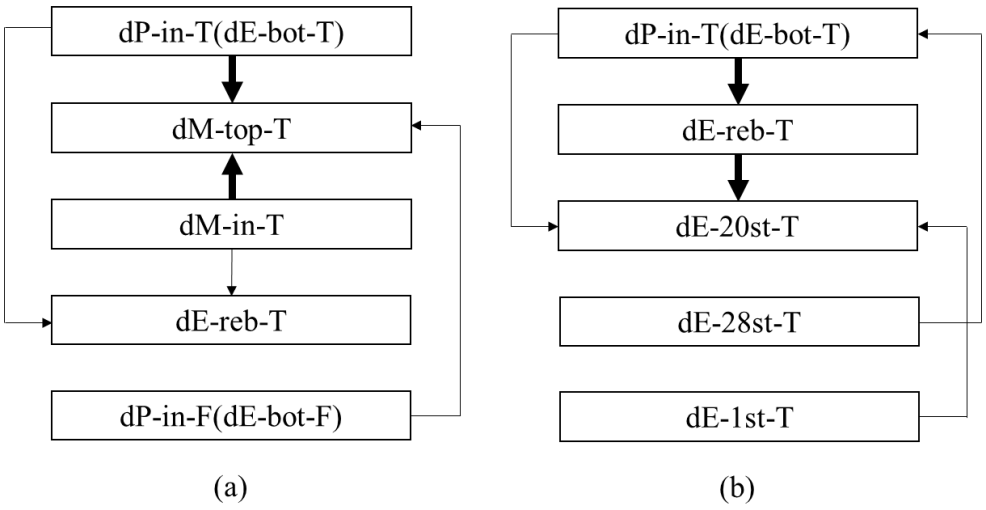


Figure 4-9 Causality flow using (a) RS amplification and (b) fault magnitude

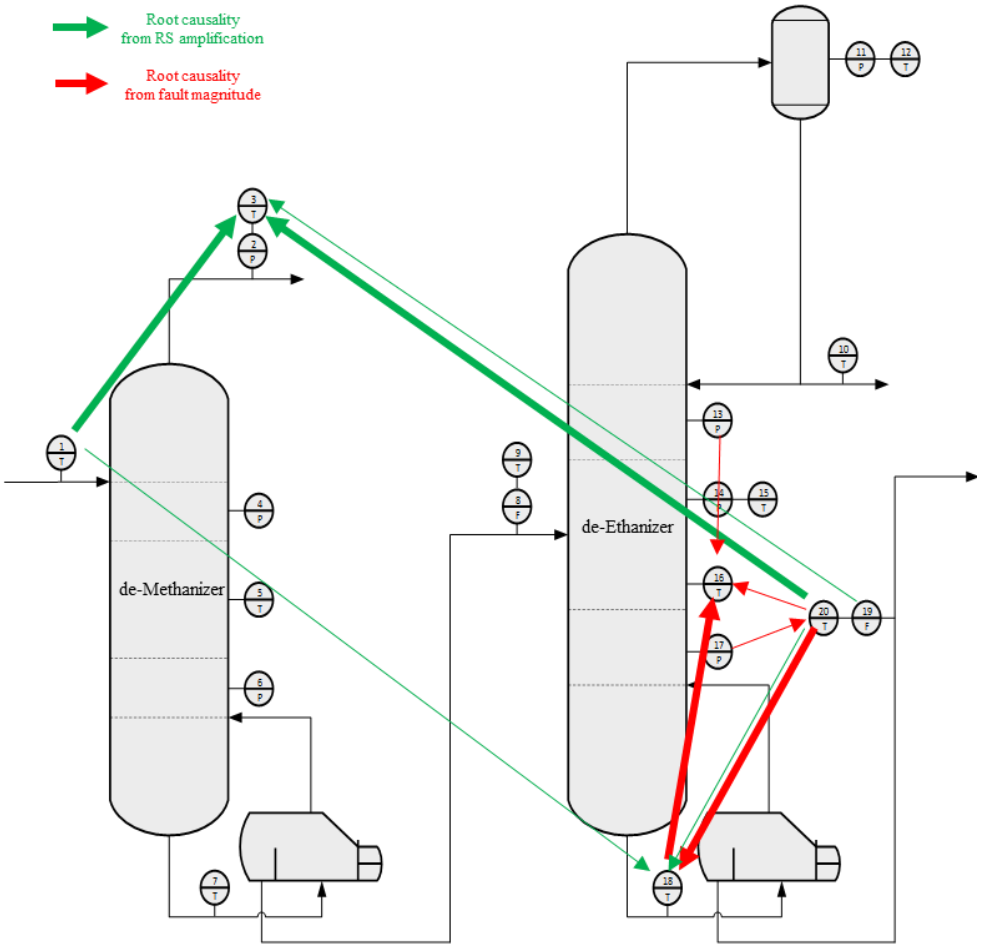


Figure 4-10 Schematic diagram the root causality from fault magnitude and RS amplification

4.3.4 **Fault scenario 2 : C-1-3%**

The fault of second scenario is deEthanizer inlet stream leaking. Leaking is very common fault in plant. To simulate this fault, 3 % inlet stream bypassed. Figure 4-11 shows the alarm in Hotelling's T^2 and Q-statistics. Comparing these 2 charts, detecting time in PCS is 436.0 seconds and alarm in RS occurs at 926.0 seconds. Because PCS detects the alarm earlier than RS does, the analysis should occur through the PCS. In the same way as fault scenario 1, to demonstrate the excellence of the algorithm, conventional PCA contribution and RS amplification are compared with the developed algorithm.

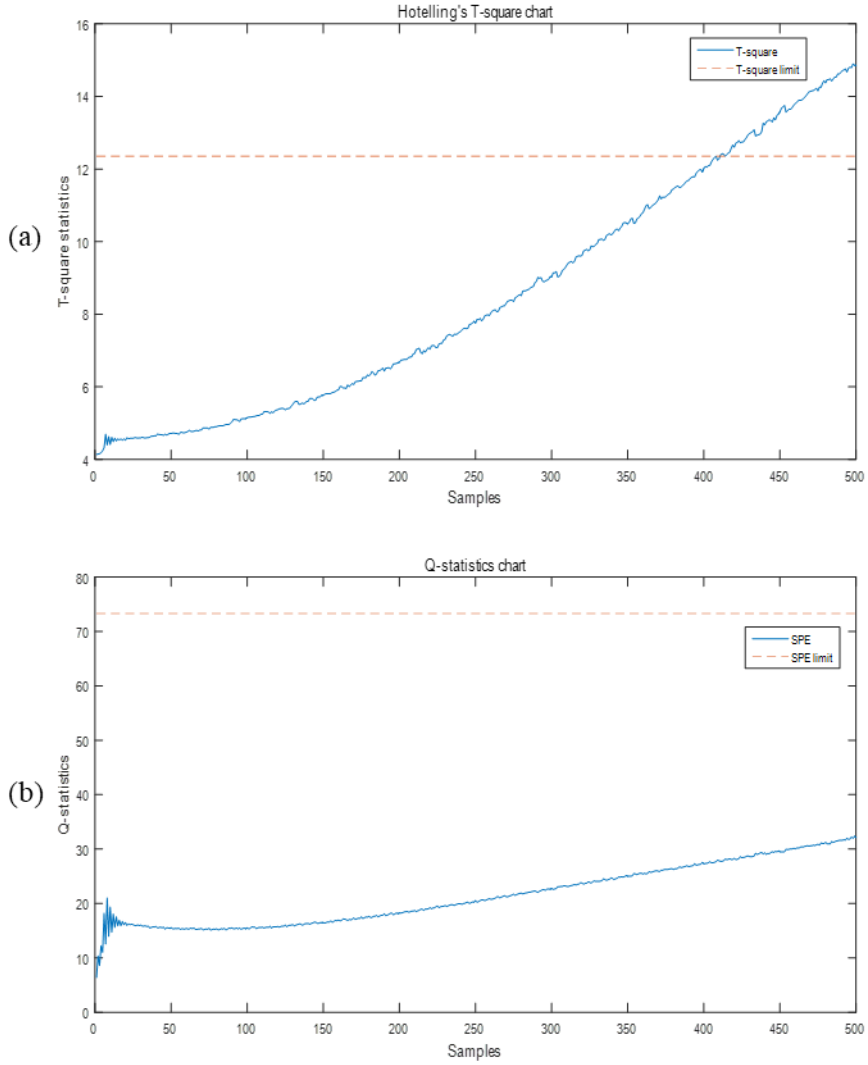


Figure 4-11 (a) Hotelling's T^2 chart (b) SPE chart at 0~500 seconds fault occurred

First, figure 4-12 shows the T-contribution chart at alarm time. This chart shows mainly pointing deEthanizer column sensors. If the contribution chart provide proper information about leaking at deEthanizer inlet stream, flow sensors of deEthanizer, such as dE-in-F (dM-bot-F) and dP-in-F (dE-bot-F), should provide larger values than those from the other sensors. However, flow sensors are not high and irrelevant sensors such as dB-in-T (dP-bot-T) and dB-reb-T are higher so that it can not used for MVGC analysis.

Using the RS amplification method, shown in figure 4-13, it makes dE-in-F (dM-bot-F) dramatically high and deEthanizer sensors are also high ranked. It seems good for finding affected sensors. However dM-in-P, dM-1st-P, dM-13st-P and dB-in-T (dP-bot-T) can be used in hierarchy sensors so that make a misleading in causality analysis.

Using developed algorithm, figure 4-14 shows the fault magnitude, which is removed normal portion. It shows deMethanizer and deButanizer sensors are decreased, so irrelevant sensors are removed. Also it makes dE-in-F (dM-bot-F) and dP-in-F (dE-bot-F) dramatically high.

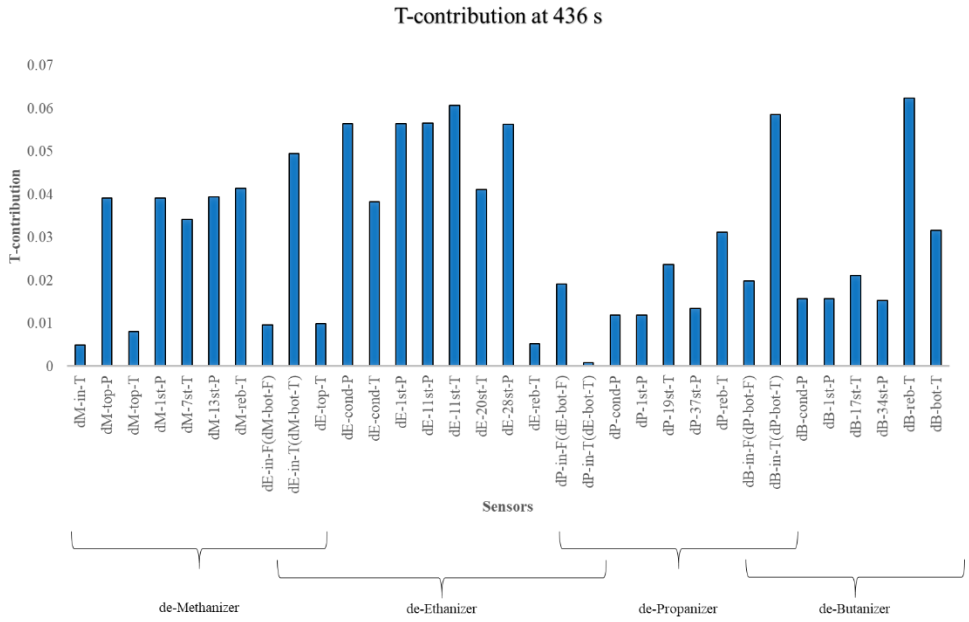


Figure 4-12 T-contribution chart at alarm occurred

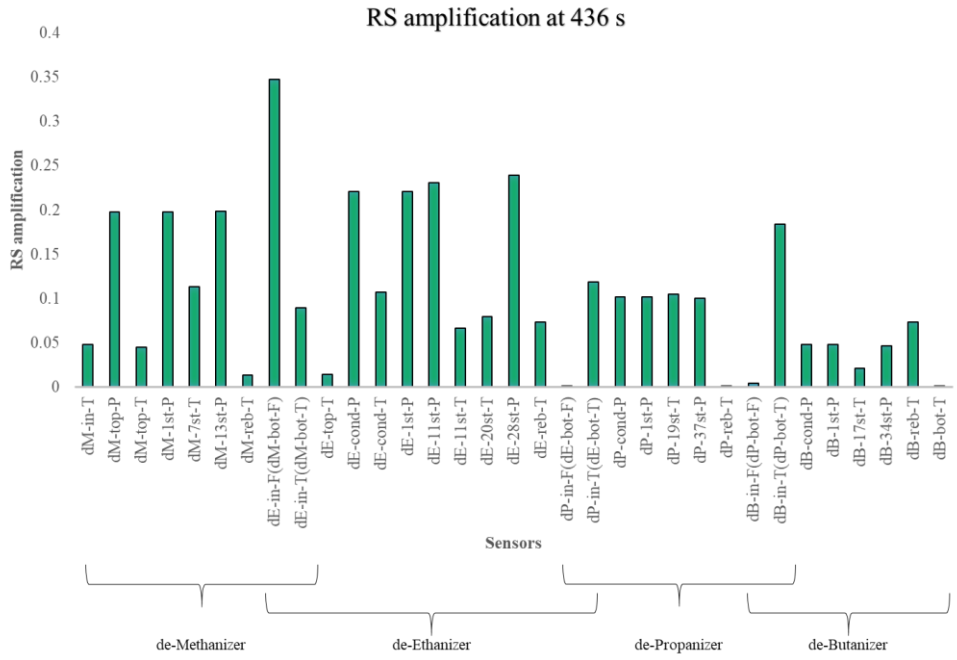


Figure 4-13 RS amplification chart at alarm occurred

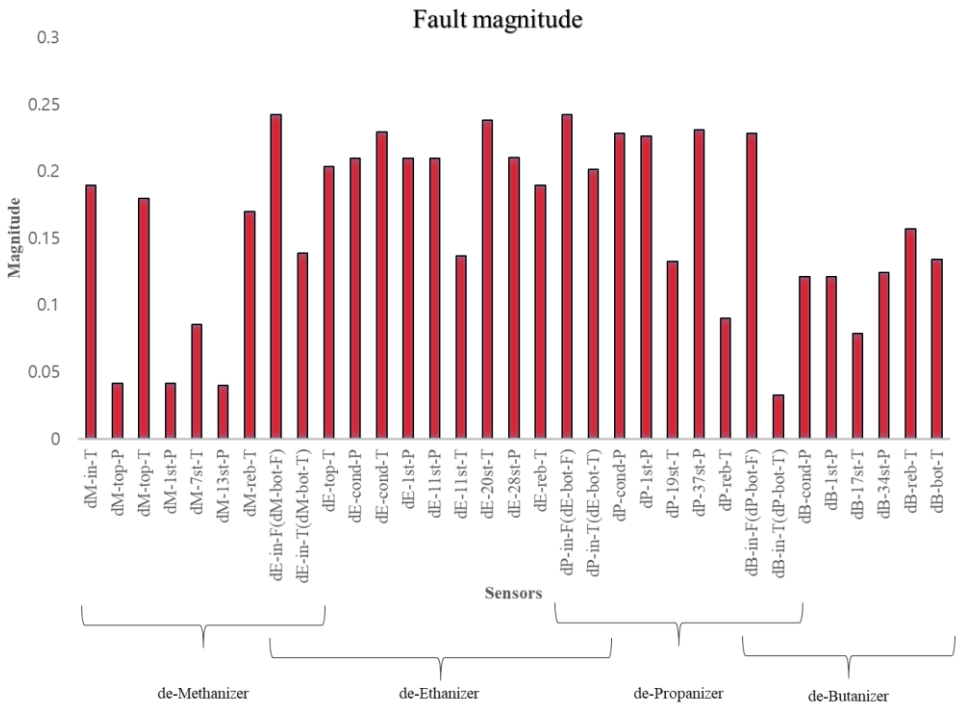


Figure 4-14 Fault magnitude chart at 337~436 seconds

Using the method mentioned in Figure 4-2, the magnitude sensors are selected for analysis in the MVGC method. From these sensors, MVGC analysis constructs the causality matrix. This matrix shows that proposed method can provide the root causality more clearly than the RS amplification method can. The causality matrix is described in Table 4-3 and Table 4-4 for the RS amplification method and fault magnitude method, respectively. The proposed method shows the greatest strength of cause and effect from dE-in-F (dM-bot-F) to dP-in-F (dE-bot-F). This causality is significantly greater than other values, so it can be recognized and interpreted as flow fault easily. On the RS amplification, the most strength cause and effect is from dB-in-T (dP-bot-T) to dM-1st-P. Second is from dE-cond-P to dB-in-T (dP-bot-T). Intuitively, these causalities are not reasonable because the physical distance between the sensors is too far or these are not sensor relationships that can affect or be affected. These compared root causalities are described in Figure 4-15. For a visual comparison, the fault magnitude method and RS amplification method are expressed in the process diagram, as shown in Figure 4-16. Green arrows indicate the RS amplification causality, and red arrows represent the fault magnitude causality. The developed method should be recognized more clearly in the process diagram.

Table 4-3 Granger causality using RS amplification method

	dE-in-F (dM-bot-F)	dE-28st-P	dE-11st-P	dE-cond-P	dM-13st-P	dM-top-P	dM-1st-P	dB-in-T (dP-bot-T)
dE-in-F (dM-bot-F)	-	0.01308	0.00000	0.00487	0.02799	0.05371	0.00417	0.00476
dE-28st-P	0.00006	-	0.00447	0.03533	0.00278	0.00075	0.00020	0.00131
dE-11st-P	0.00184	0.00002	-	0.01042	0.00098	0.00027	0.00217	0.00462
dE-cond-P	0.00138	0.00017	0.00142	-	0.00024	0.00165	0.00118	0.00124
dM-13st-P	0.00033	0.00023	0.00054	0.00628	-	0.00066	0.00444	0.00391
dM-top-P	0.00197	0.00050	0.00083	0.00080	0.00158	-	0.00106	0.00015
dM-1st-P	0.00100	0.00335	0.01288	0.03507	0.00084	0.00389	-	0.13892
dB-in-T (dP-bot-T)	0.00003	0.00042	0.00635	0.07209	0.05685	0.00005	0.00173	-

Table 4-4 Granger causality using proposed method

	dE-in-F (dM-bot-F)	dP-in-F (dE-bot-F)	dE-20st-T	dP-37st-P	dE-cond-T	dB-in-F (dP-bot-F)	dP-cond-P
dE-in-F (dM-bot-F)	-	0.0226	0.0025	0.0157	0.0024	0.0038	0.0035
dP-in-F (dE-bot-F)	0.5183	-	0.0614	0.0263	0.0409	0.0039	0.0031
dE-20st-T	0.0559	0.0778	-	0.0220	0.0043	0.0019	0.0000
dP-37st-P	0.0146	0.0191	0.0841	-	0.0000	0.0078	0.0000
dE-cond-T	0.0922	0.0059	0.0173	0.0156	-	0.0236	0.0007
dB-in-F (dP-bot-F)	0.0250	0.0205	0.0658	0.0048	0.0001	-	0.0046
dP-cond-P	0.0002	0.0659	0.0830	0.0120	0.0192	0.0066	-

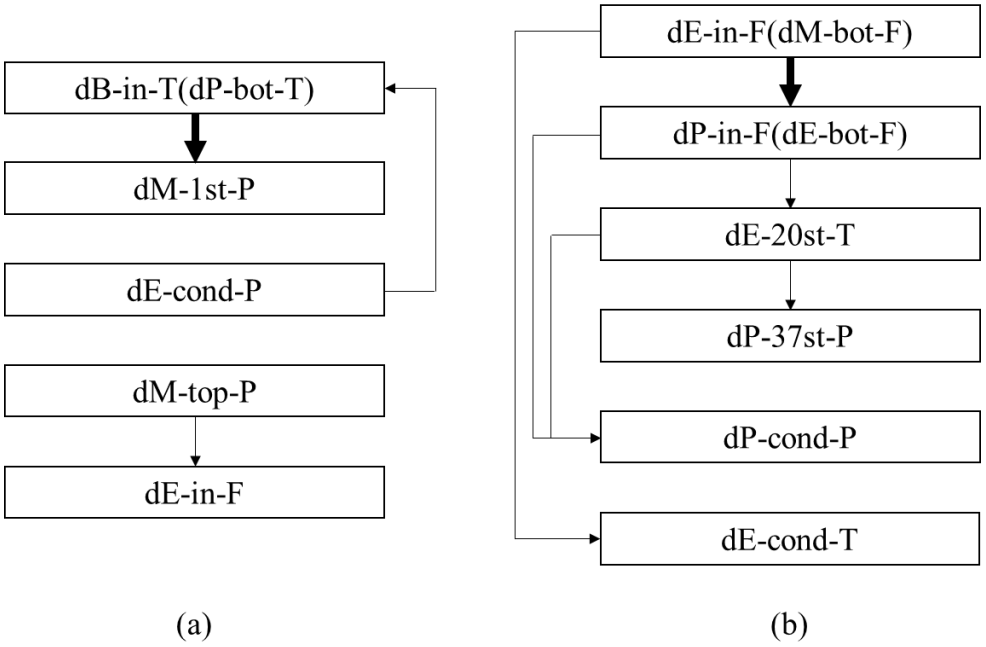


Figure 4-15 Causality flow using (a) RS amplification and (b) fault magnitude

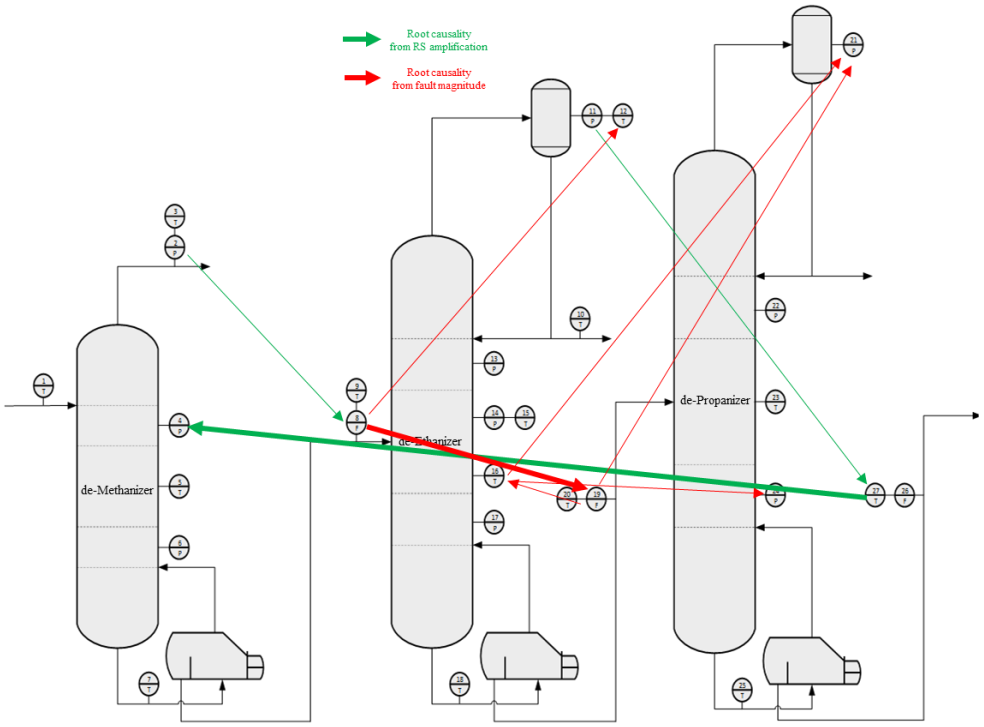


Figure 4-16 Schematic diagram the root causality from fault magnitude and RS amplification

4.3.5 **Fault scenario 3 : A-4-3%**

Final scenario is condenser overcooling in deEthanizer column. It is also common fault in deEthanizer. In this case, compared with the scenario 1, it is expected that propagation path is to be opposite direction. For simulating this fault, heat duty of deEthanizer condenser is increased 3 % than normal state. In this case, SPE alarm is occurred earlier than Hotelling's T^2limit. The detecting time is 165.0 seconds in RS, and not alarmed until about 2000 seconds in PCS. Figure 4-17 (a) and (b) show these result. Therefore, the root causality should be analyzed in RS. To prove the excellence of the algorithm, the conventional PCA contribution and RS amplification are compared with the developed algorithm.

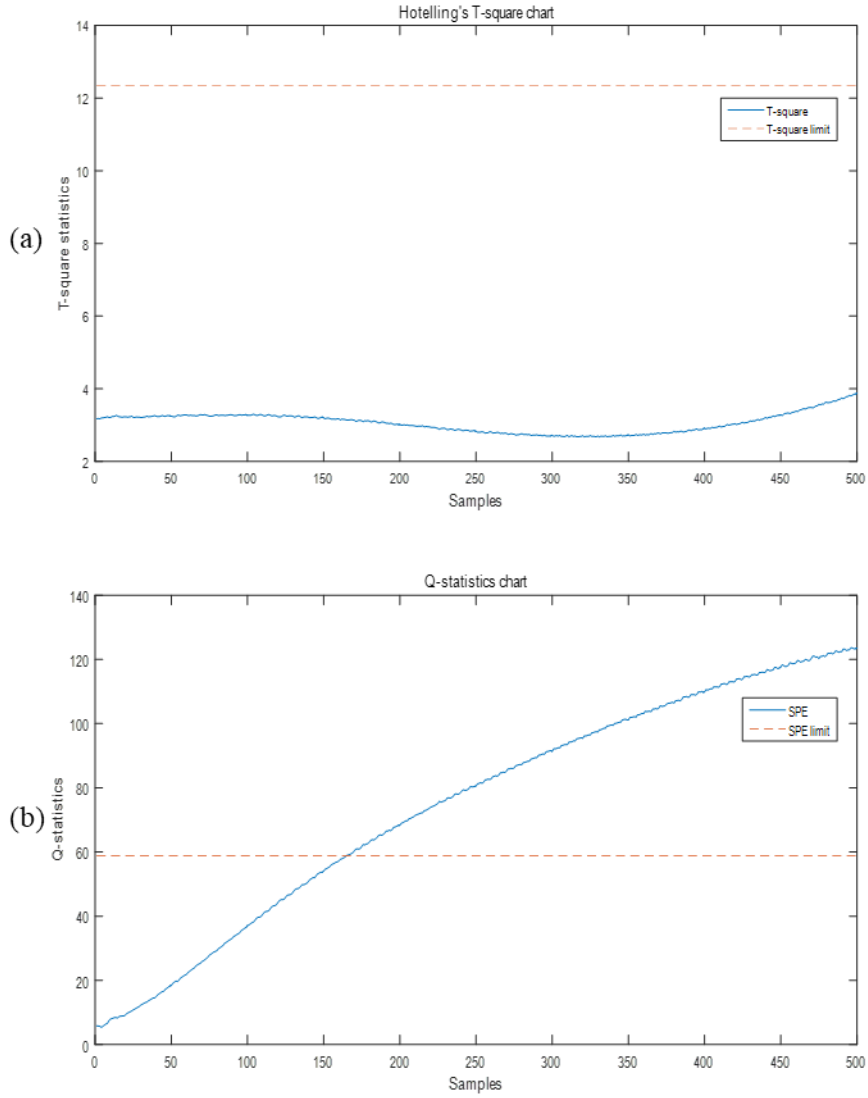


Figure 4-17 (a) Hotelling's T^2 chart (b) SPE chart at 0~500 seconds fault occurred

Using the SPE contribution at 165 s, Figure 4-18 shows the SPE contribution chart at the alarm time. Sensors dE-cond-P, dE-1st-P, dE-11st-P, dE-28st-P, and dP-in-F (dE-bot-F) are the most affected. Using RS amplification makes this result clearer. Figure 4-19 shows that dE-cond-P, dE-1st-P, dE-11st-P, dE-28st-P, and dP-19st-T are enhanced, and the other sensors are weaken by the RS amplification method. This is a suitable result because these sensors are related to the condenser. The proposed method, fault magnitude, shows that dE-top-T, dE-cond-P, dE-1st-P, dE-11st-P, and dE-28st-P are the major fault sensors, as shown in Figure 4-20. These results appear to be similar except for one or two sensors; dP-in-F (dE-bot-F) in SPE-contribution, dP-19st-T in RS-amplification, and dE-top-T and dE-1st-P in the fault magnitude method.

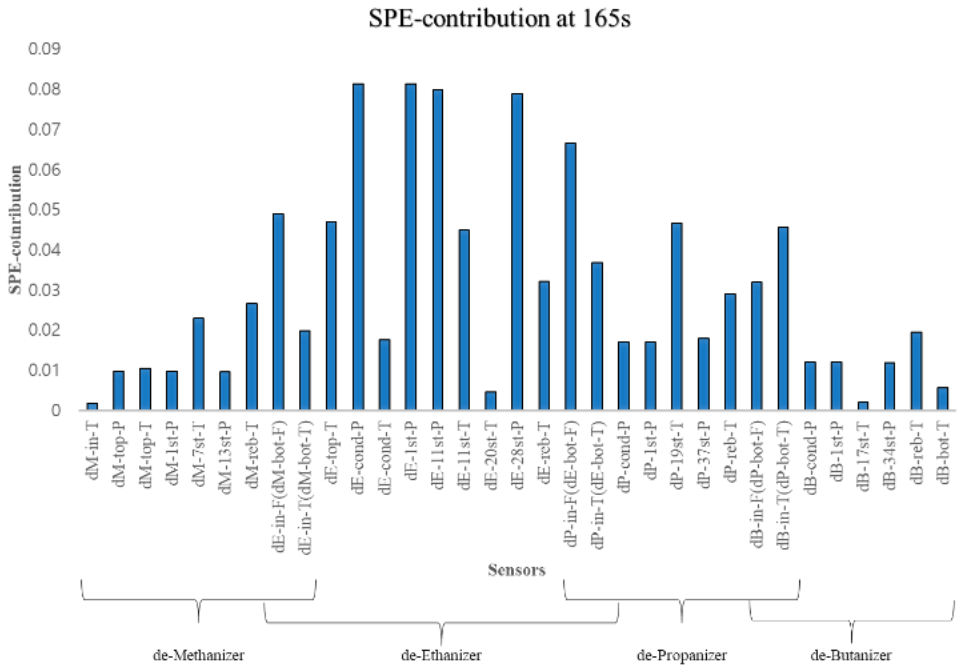


Figure 4-18 SPE contribution at alarm occurred

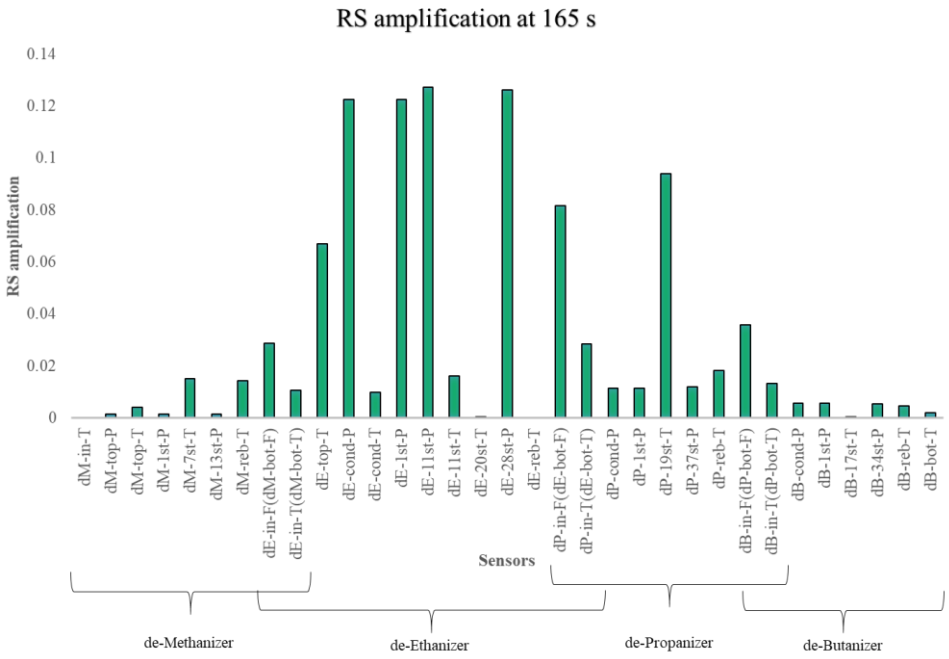


Figure 4-19 RS amplification chart at alarm occurred

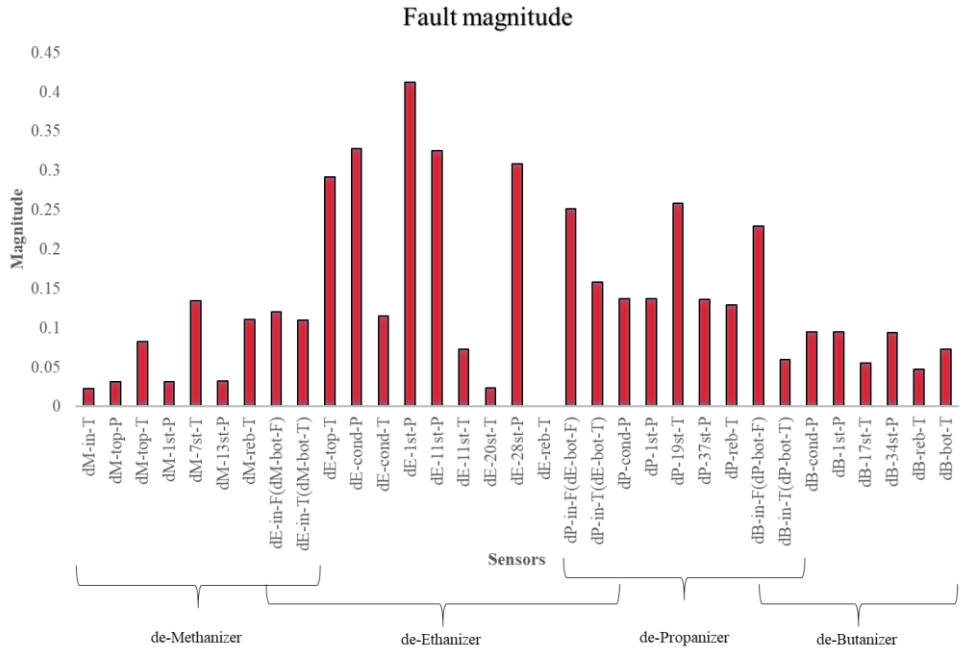


Figure 4-20 Fault magnitude chart at 66~165 seconds

A small difference in the result of the three methodologies, produces a completely different result in the MVGC analysis. Tables 4-5, shows the causality from the three methods, SPE-contribution, RS-amplification, and fault magnitude, respectively.

First, the SPE contribution shows that dE-cond-P affects dP-in-F (dE-bot-F), dE-28st-P affects dE-cond-P and dE-11st-P, and dP-in-F (dE-bot-F) affects dE-cond-P and dE-28st-P. These causalities can be interpreted that the relationship between causality sensors is a bit far, broad, and crossed, such that it is ambiguous to define the root cause. RS amplification provides the main causalities as dE-11st-P to dE-28st-P and dE-cond-P. The weak causalities are from dE-28st-P to dE-11st-P, from dE-cond-P to dE-11st-P, and from dP-19st-T to dE-28st-P. These results can be interpreted as the fault starts from the column internal pressure problem. Finally, the fault magnitude methodology shows that dE-top-T affects dE-1st-P primarily, and weak causalities are given by dE-cond-P to dE-1st-P and dE-top-T. This result indicates that the root cause starts from the condenser area. These results are described as a flow diagram in Figure 4-21 and can be more clearly visualized in the process diagram in Figure 4-22. This result comes from the difference of just two sensors, compared with the SPE contribution and RS amplification. It can be interpreted that causality must include all the major variables about the fault. If one key variable is missing, the result can be misleading, as results show. Therefore, the fault magnitude algorithm can select hierarchical sensors properly and find the accurate root cause at the initial stage of fault.

Table 4-5 Granger causality using proposed method

	dE-1st-P	**dE-cond-P**	**dE-11st-P**	**dE-28st-P**	**dE-top-T**
dE-1st-P	-	0.021268	0.003558	0.001124	0.092738
dE-cond-P	0.00171	-	0.002833	6.89E-08	0.000581
dE-11st-P	0.000144	0.001618	-	0.002808	5.69E-05
dE-28st-P	0.000658	0.001038	0.003933	-	0.00041
dE-top-T	0.003782	0.017225	0.002918	1.02E-04	-

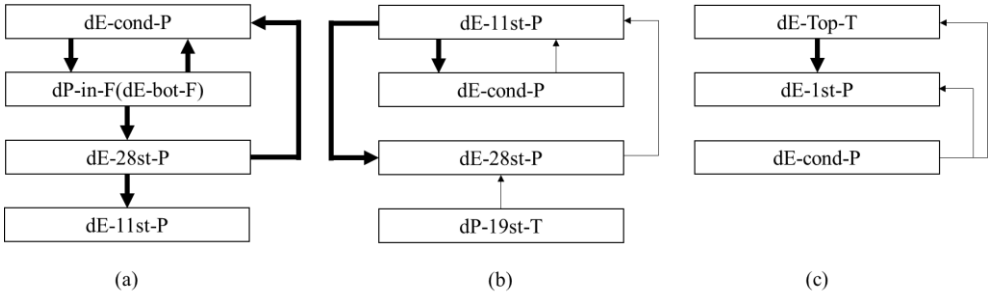


Figure 4-21 Causality flow using (a) SPE contribution (b) RS amplification and (c) fault magnitude

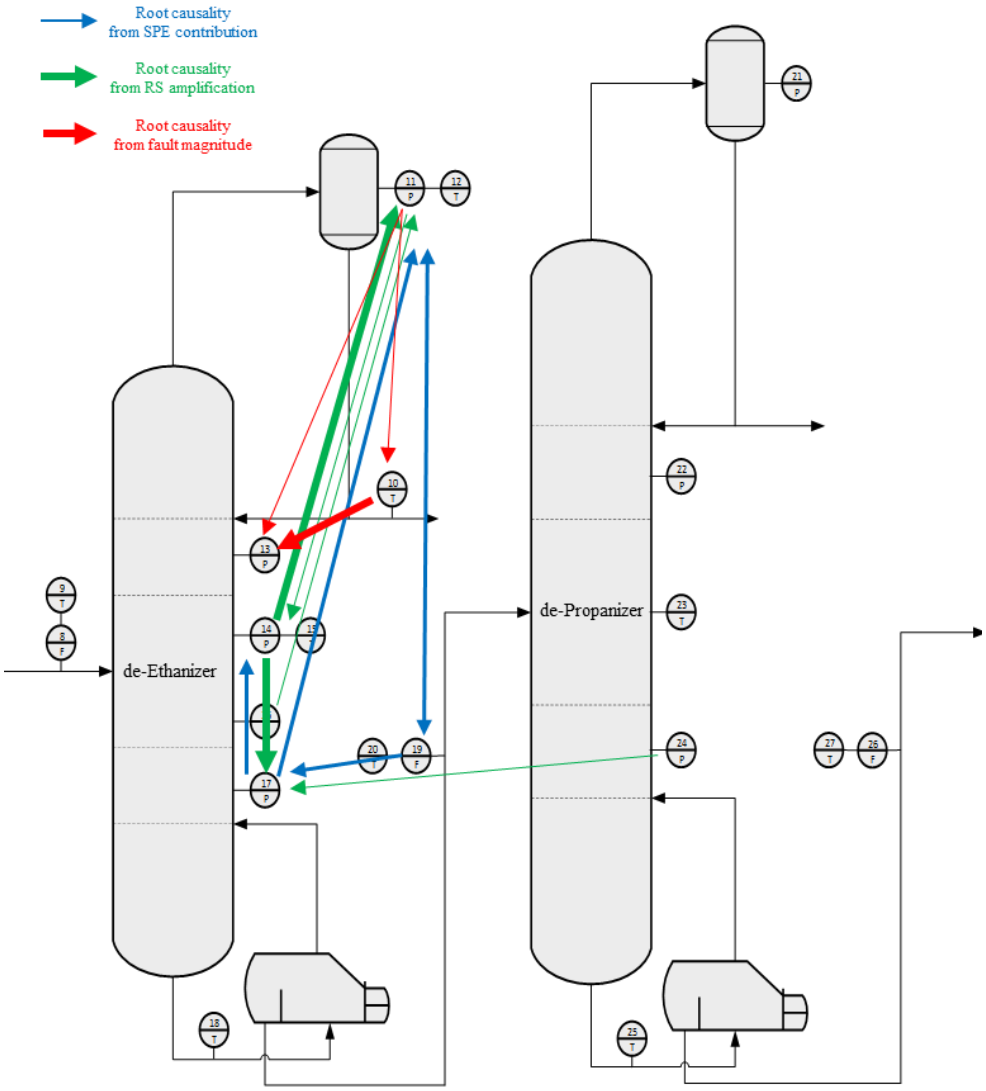


Figure 4-22 Schematic diagram the root causality from fault magnitude, RS amplification and SPE contribution

4.4 Conclusion

This study proposes a fault analysis using divide subspace for minimize the normal portion in fault information. First, PCS and RS respectively used for proper removing of normal portion. Scaled fault contribution decomposed by SVD and from the fault magnitude hierarchy variables are selected. MVGC calculated these major variables so that make a matrix of causality. For verifying the performance, root cause compared with conventional contribution and RS amplification. Because the fault scale is too small, contribution information at initial stage can not provide root cause properly. RS amplification is good performance in enhancement contribution affected variables. However, in terms of the MVGC analysis, RS amplification provide misleading when the alarm occurred in PCS. Also, because the key variables are essential in MVGC analysis, developed method is able to obtain better performance than RS amplification in root causality analysis. Proposed method uses both space, PCS and RS according to criteria, all the key variables that reflect the process state, it can make an accurate root cause.

CHAPTER 5 **Concluding Remarks**

In this thesis, a multi-mode process monitoring for early detection and robust root cause diagnosis for initial fault stage are proposed. The methodology is consist of 2 parts; first part is normal operation modeling from global to local based on PCA and classification. Second part is monitoring and fault detection based on MDM, kNN and PCA and final part is root cause diagnosis based on PCA, SVD and MVGC.

First, process overall normal operation data is decomposed into reduced space from PCA method. T-score, which is calculated with reducing dimensions, is used for k-means clustering method. The value of k is defined intuitively using the T-score chart. All normal training variables have own group class and PCA makes new local model using variables in group. This is defined local PCA modeling. After the modeling, each mode makes own process limitation, Hotelling's T^2 and SPE, it is ready ready for monitoring and fault detection.

Secondly, a new sample variable is projected into global PCA model. From T-score and MDM, kNN methods, it can be assigned class with training local normal data. In assigned group, it is determined fault or normal. When the fault occurred, it goes through root cause diagnosis part.

Finally, fault data is determined PCS or RS from alarm index. If it beyond Hotelling's T^2 alarm limit, it is calculated in PCS, or beyond SPE alarm, it is

calculated in RS. In each space, contribution plot is scaled from normal contribution, and then it is decomposed by SVD. From the SVD, fault magnitude is derived and also empirical rule from statistics select the hierarchy sensors. These sensors are analyzed by MVGC and the result is root casual information.

To verify the proposed methodology, LNG plant fractionation process is applied. A total of 45 case studies is used for comparing. In monitoring result, proposed method is compare with global PCA and univariate method, shewart 3-sigma based on FDR and FDA. First, proposed method has only 2 cases are drawing and all the rest are exceedingly better performance than global PCA. Compared with univariate method, 35 cases are enough better performance, 8 cases are the same and only 2 cases are a little poor but there is few difference. Proposed method has the average of FDR value is 96.2 where univariate is 91.4 and global PCA is 85.5. In FDA, 97.0 is proposed method, 93.1 is univariate method and 88.4 is global PCA. From these result, proposed model increase the monitoring accuracy and detection rate.

In diagnosis part, root cause from proposed method have good performance. To verify the performance, conventional contribution chart at fault time and RS amplification method are compared with proposed method. From the graph chart, proposed method can isolate normal portion from fault data. Therefore, only proposed method can provide proper root cause at fault initial detection stage. Especially, initial alarm propagation is very similar with proposed root cause. This method is only used normal historic data and assumed small intensity of fault, it can be adjusted most plant.

Future work is to enhance the classification part. Because using the dynamic model, it has limitation of various normal operation mode so it can be perfectly performed the classification. Lastly, when the fault occurred, in this study 100 seconds fault data are used, but there is no criteria. If sensitivity analysis and decision rule is defined, the robustness of this method can be increased.

Bibliography

(1) He, Q. P.; Qin, S. J.; Wang, J. A New Fault Diagnosis Method Using Fault Directions in Fisher Discriminant Analysis. *AIChE Journal* **2005**, *51* (2), 555–571. https://doi.org/10.1002/aic.10325.

(2) Alzate, C.; Suykens, J. A. K. Multiway Spectral Clustering with Out-of-Sample Extensions through Weighted Kernel PCA. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2010**, *32* (2), 335–347. https://doi.org/10.1109/TPAMI.2008.292.

(3) Nomikos, P.; MacGregor, J. F. Monitoring Batch Processes Using Multiway Principal Component Analysis. *AIChE Journal* **1994**, *40* (8), 1361–1375. https://doi.org/10.1002/aic.690400809.

(4) Stubbs, S.; Zhang, J.; Morris, J. Multiway Interval Partial Least Squares for Batch Process Performance Monitoring. *Industrial & Engineering Chemistry Research* **2013**, *52* (35), 12399–12407. https://doi.org/10.1021/ie303562t.

(5) Rongfu Luo; Manish Misra, and; Himmelblau*, D. M. Sensor Fault Detection via Multiscale Analysis and Dynamic PCA. **1999**. https://doi.org/10.1021/IE980557B.

(6) Misra, M.; Yue, H. H.; Qin, S. J.; Ling, C. Multivariate Process Monitoring and Fault Diagnosis by Multi-Scale PCA. *Computers and Chemical Engineering* **2002**, *26* (9), 1281–1293. https://doi.org/10.1016/S0098-1354(02)00093-5.

(7) Hou, Z.-S.; Wang, Z. From Model-Based Control to Data-Driven Control: Survey, Classification and Perspective. *Information Sciences* **2013**, *235*, 3–35. https://doi.org/10.1016/J.INS.2012.07.014.

(8) Hosseini, A. H.; Hussain, S.; Gabbar, H. A. Simulation-Based Fault Propagation Analysis - Application on Hydrogen Production Plant. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2013**, *7906 LNAI* (6), 441–448. https://doi.org/10.1007/978-3-642-38577-3_45.

(9) Xu, R.; WunschII, D. Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks* **2005**, *16* (3), 645–678. https://doi.org/10.1109/TNN.2005.845141.

(10) Geng, Z.; Zhu, Q. Multiscale Nonlinear Principal Component Analysis (NLPCA) and Its Application for Chemical Process Monitoring. *Industrial*

and Engineering Chemistry Research **2005**, *44* (10), 3585–3593. https://doi.org/10.1021/ie0493107.

(11) Yoon, S.; MacGregor, J. F. Fault Diagnosis with Multivariate Statistical Models Part I: Using Steady State Fault Signatures. *Journal of Process Control* **2001**, *11* (4), 387–400. https://doi.org/10.1016/S0959-1524(00)00008-1.

(12) Gowid, S.; Dixon, R.; Ghani, S. Profitability, Reliability and Condition Based Monitoring of LNG Floating Platforms: A Review. *Journal of Natural Gas Science and Engineering* **2015**, *27*, 1495–1511. https://doi.org/10.1016/j.jngse.2015.10.015.

(13) MacGregor, J. F.; Kourti, T. Statistical Process Control of Multivariate Processes. *Control Engineering Practice* **1995**, *3* (3), 403–414. https://doi.org/10.1016/0967-0661(95)00014-L.

(14) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta* **1986**, *185*, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.

(15) Choi, S. W.; Lee, C.; Lee, J.-M.; Park, J. H.; Lee, I.-B. Fault Detection and Identification of Nonlinear Processes Based on Kernel PCA. *Chemometrics and Intelligent Laboratory Systems* **2005**, *75* (1), 55–67. https://doi.org/10.1016/J.CHEMOLAB.2004.05.001.

(16) Ge, Z.; Song, Z. Two-Level Multiblock Statistical Monitoring for Plant-Wide Processes. *Korean Journal of Chemical Engineering* **2009**, *26* (6), 1467–1475. https://doi.org/10.1007/s11814-009-0283-7.

(17) Ding, M.; Chen, Y.; Bressler, S. L. Granger Causality: Basic Theory and Application to Neuroscience. In *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*; 2006; pp 437–460. https://doi.org/10.1002/9783527609970.ch17.

(18) Dai, X.; Gao, Z. From Model, Signal to Knowledge: A Data-Driven Perspective of Fault Detection and Diagnosis. *IEEE Transactions on Industrial Informatics* **2013**, *9* (4), 2226–2238. https://doi.org/10.1109/TII.2013.2243743.

(19) MacGregor, J. F.; Jaeckle, C.; Kiparissides, C.; Koutoudi, M. Process Monitoring and Diagnosis by Multiblock PLS Methods. *AIChE Journal* **1994**, *40* (5), 826–838. https://doi.org/10.1002/aic.690400509.

(20) Yue, H. H.; Qin, S. J. Reconstruction-Based Fault Identification Using a Combined Index. *Industrial and Engineering Chemistry Research* **2001**, *40*

(20), 4403–4414. https://doi.org/10.1021/ie000141+.

(21) Wen, X.; Rangarajan, G.; Ding, M. Multivariate Granger Causality: An Estimation Framework Based on Factorization of the Spectral Density Matrix. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2013**, *371* (1997). https://doi.org/10.1098/rsta.2011.0610.

(22) Chen, H. S.; Yan, Z.; Zhang, X.; Liu, Y.; Yao, Y. Root Cause Diagnosis of Process Faults Using Conditional Granger Causality Analysis and Maximum Spanning Tree. *IFAC-PapersOnLine* **2018**, *51* (18), 381–386. https://doi.org/10.1016/j.ifacol.2018.09.330.

(23) Barnett, L.; Seth, A. K. The MVGC Multivariate Granger Causality Toolbox: A New Approach to Granger-Causal Inference. *Journal of Neuroscience Methods* **2014**, *223*, 50–68. https://doi.org/10.1016/J.JNEUMETH.2013.10.018.

(24) Yin, S.; Li, X.; Gao, H.; Kaynak, O. Data-Based Techniques Focused on Modern Industry: An Overview. *IEEE Transactions on Industrial Electronics* **2015**, *62* (1), 657–667. https://doi.org/10.1109/TIE.2014.2308133.

(25) Lane, S.; Martin, E. B.; Morris, A. J.; Gower, P. Application of Exponentially Weighted Principal Component Analysis for the Monitoring of a Polymer Film Manufacturing Process. *Transactions of the Institute of Measurement and Control* **2003**, *25* (1), 17–35. https://doi.org/10.1191/0142331203tm071oa.

(26) Li, W.; Yue, H. H.; Valle-Cervantes, S.; Qin, S. J. Recursive PCA for Adaptive Process Monitoring. *Journal of Process Control* **2000**, *10* (5), 471–486. https://doi.org/10.1016/S0959-1524(00)00022-6.

(27) Garcia, M.; Ruiz, M.; Colomer, J.; Melendez, J. Multiway Principal Component Analysis and Case Base Reasoning Methodology for Abnormal Situation Detection in a Nutrient Removing SBR. In *2007 European Control Conference (ECC)*; IEEE, 2007; pp 5354–5360. https://doi.org/10.23919/ECC.2007.7068413.

(28) Zhang, Y.; Vaculik, V.; Dudzic, M.; Miletic, I.; Smyth, A.; Holek, T. Start Cast Breakouts Preventative Prediction Using Multi-Way PCA Technology. *IFAC Proceedings Volumes* **2003**, *36* (24), 101–106. https://doi.org/10.1016/S1474-6670(17)37611-5.

(29) Westerhuis, J. A.; Kourti, T.; MacGregor, J. F. Analysis of Multiblock and Hierarchical PCA and PLS Models. *Journal of Chemometrics* **1998**, *12* (5),

301–321. https://doi.org/10.1002/(SICI)1099-128X(199809/10)12:5<301::AID-CEM515>3.0.CO;2-S.

(30) Smilde, A. K.; Westerhuis, J. A.; de Jong, S. A Framework for Sequential Multiblock Component Methods. *Journal of Chemometrics* **2003**, *17* (6), 323–337. https://doi.org/10.1002/cem.811.

(31) Bahadori, A. *Natural Gas Processing: Technology and Engineering Design*; Elsevier, 2014. https://doi.org/10.1016/C2013-0-13070-X.

(32) Bakshi, B. R. Multiscale PCA with Application to Multivariate Statistical Process Monitoring. *AIChE Journal* **1998**, *44* (7), 1596–1610. https://doi.org/10.1002/aic.690440712.

(33) Dudzic, M.; Vaculik, V.; Miletic, I. Applications of Multivariate Statistics at Dofasco. In *IEEE Industry Applications Society Advanced Process Control Applications for Industry Workshop*; IEEE; pp 27–29. https://doi.org/10.1109/APCA.1999.805022.

(34) Bazin, C.; Hodouin, D.; Desbiens, A.; Canadian Institute of Mining, M. and P.; Conference of Metallurgists (38th : 1999 : Québec, Q. *Control and Optimization in Minerals, Metals and Materials Processing*; Canadian Institute of Mining, Metallurgy and Petroleum, 1999.

(35) Dudzic, M.; Vaculik, V.; Miletic, I. On-Line Applications of Multivariate Statistics at Dofasco. *IFAC Proceedings Volumes* **2000**, *33* (22), 425–430. https://doi.org/10.1016/S1474-6670(17)37032-5.

(36) Jeong, H.; Cho, S.; Kim, D.; Pyun, H.; Ha, D.; Han, C.; Kang, M.; Jeong, M.; Lee, S. A Heuristic Method of Variable Selection Based on Principal Component Analysis and Factor Analysis for Monitoring in a 300 KW MCFC Power Plant. *International Journal of Hydrogen Energy* **2012**, *37* (15), 11394–11400. https://doi.org/10.1016/J.IJHYDENE.2012.04.135.

(37) Jiang, Q.; Yan, X. Monitoring Multi-Mode Plant-Wide Processes by Using Mutual Information-Based Multi-Block PCA, Joint Probability, and Bayesian Inference. *Chemometrics and Intelligent Laboratory Systems* **2014**, *136*, 121–137. https://doi.org/10.1016/j.chemolab.2014.05.012.

(38) Venkatasubramanian, V.; Rengaswamy, R.; Kavuri, S. N. A Review of Process Fault Detection and Diagnosis Part II: Qualitative Models and Search Strategies. *Computers and Chemical Engineering* **2003**, *27* (3), 313–326. https://doi.org/10.1016/S0098-1354(02)00161-8.

(39) Venkatasubramanian, V.; Rengaswamy, R.; Kavuri, S. N.; Yin, K. A Review of Process Fault Detection and Diagnosis. *Computers & Chemical*

Engineering **2003**, *27* (3), 327–346. https://doi.org/10.1016/s0098-1354(02)00162-x.

(40) Qin, S. J. Survey on Data-Driven Industrial Process Monitoring and Diagnosis. *Annual Reviews in Control* **2012**, *36* (2), 220–234. https://doi.org/10.1016/j.arcontrol.2012.09.004.

(41) Dvorak, D.; Kuipers, B. Process Monitoring and Diagnosis: A Model-Based Approach. *IEEE Expert-Intelligent Systems and their Applications* **1991**, *6* (3), 67–74. https://doi.org/10.1109/64.87688.

(42) Narozny, J. A.; Phohole, A. D. L.; Elnir, M. An Expert System for Machine Fault Diagnosis. *IFAC Proceedings Volumes* **2001**, *34* (18), 95–99. https://doi.org/10.1016/S1474-6670(17)33188-9.

(43) Joe Qin, S. Statistical Process Monitoring: Basics and Beyond. *Journal of Chemometrics* **2003**, *17* (8–9), 480–502. https://doi.org/10.1002/cem.800.

(44) Ahmed, U.; Ha, D.; An, J.; Zahid, U.; Han, C. Fault Propagation Path Estimation in NGL Fractionation Process Using Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **2017**, *162* (January), 73–82. https://doi.org/10.1016/j.chemolab.2017.01.006.

(45) Ahmed, U.; Ha, D.; Shin, S.; Shaukat, N.; Zahid, U.; Han, C. Estimation of Disturbance Propagation Path Using Principal Component Analysis (PCA) and Multivariate Granger Causality (MVGC) Techniques. *Industrial and Engineering Chemistry Research* **2017**, *56* (25), 7260–7272. https://doi.org/10.1021/acs.iecr.6b02763.

(46) Kitano, K.; Kano, M.; Gopaluni, B. Fault Identification with Modified Reconstruction-Based Contribution Based on Kernel Principal Component Analysis. In *2017 Asian Control Conference, ASCC 2017*; IEEE, 2018; Vol. 2018-Janua, pp 1493–1498. https://doi.org/10.1109/ASCC.2017.8287394.

(47) Stichlmair, J.; Fair, J. R. *Distillation : Principles and Practices*; Wiley, 1998.

Nomenclature and Abbreviations

FDA : Fault detection accuracy

FDR : Fault detection rate

kNN : k-nearest neighbors

GC: Granger Causality

LNG : Liquefied natural gas

MDM : Minimum distance to mean

MVGC : Multivariate granger causality

PCA : Principal component analysis

PCS : Principal component subspace

RS : Residual subspace

SPE : Squared prediction error

SVD : Singular value decomposition

Abstract in Korean (요 약)

데이터 저장, 처리 속도가 발전하면서 공정 데이터 분석 영역 또한 지난 수십년 동안 급속도로 발전하였다. 그 결과 많은 공장들이 단변량 뿐 아니라 다변량 통계 기법을 활용하여 실시간으로 빠른 이상 감시를 이루어 내고 있다. 지속적으로 축적되는 정상 데이터를 통해 이상과 정상을 구분하는 방법은 점점 더 빠르고 정확해 지고 있다.

하지만 이상 진단 영역은 빠른 이상 감시와는 달리 많은 제약들이 존재하고 있다. 공정 이상을 진단하는 방법은 정성적인 모델 분석 방법, 전문가 시스템과 같은 지식을 기반으로 하는 방법, 공정 감시와 같이 데이터를 기반으로 하는 방법으로 나누어 진다. 이 중 정성적인 모델 분석 방법은 공정이 커지고 복잡해 지면서 모든 이상 상황에 적합한 대응 정보를 제공하는 것은 현실적으로 불가능하다. 전문가 시스템과 같이 지식을 기반으로 한 이상 진단은 그 정확도는 높을 수 있으나, 분석 시간이 오래 걸리기 때문에 주로 사고 후 진단에 활용되는 것이 일반적이다. 이러한 제약들 때문에 실시간으로 이상을 진단하는 방법론들은 주로 과거 데이터를 기반으로 한 분석 방법이다. 하지만 과거 데이터를 기반으로 한 대다수의 실시간 이상 진단 방법론들은 실제 일어났던 이상 자체를 분석하여 그 특정 이상에 최적화된 관리 방법을 제공해 주는 방법을 사용하고 있기에 그 적용 범위가 협소하고 공정 상태에 따라 그 정확도 편

차도 심각하게 달라질 수 있다.

실시간 이상 진단이 갖는 이러한 어려움을 해결하고자 본 논문에서는 빠르게 감지되는 이상 시간과 동시에 그 이상의 근원 정보를 제공하는 연구를 수행하였다. 특히, 이상 감지 시간에 그 이상 원인을 찾기 힘든, 이상의 크기가 매우 작은 경우를 가정하고 연구를 진행하였다.

첫째, 이상을 빠르게 감시하기 위해 다변량 통계 기법 중 가장 기본적인 주성분 분석 방법론을 사용하였으며, 정확도 및 감시 속도 성능을 높이기 추가적인 모델링 과정을 거쳤다. 먼저 주성분 분석 방법론에서 도출되는 Hotelling's T^2 값을 k-평균 군집법으로 군집화 하여, 여러 개의 정상 운전 모드를 나누었다. 나누어 진 정상 운전 모드는 각각 다시 주성분 분석 방법으로 모델링 하여 개별적인 로칼 주성분 분석 모델을 도출하였다.

둘째, 여러 개의 로칼 주성분 분석 모델을 효과적으로 매칭하여 이상을 감시하기 위해, 최소거리평균법과 k-최근접 이웃 알고리즘을 적용하였다. 본 방법론을 적용하여 실시간 데이터를 매칭된 로칼 주성분 분석으로 이상을 감시하였다

마지막으로, 실시간 이상 진단 정보 제공을 위해 주성분 분석의 컨트리뷰션 정보와 특이값 분해, 다변수 그레인져 인과관계 방법론을 사용하였다. 이상의 크기가 작은 이상들을 대상으로 하기 때문에, 이상 데이터에

섞여 있는 정상 정보를 제거해 주는 방법이 필요하다. 본 문제를 해결하기 위해 주성분 분석 후 이상에 기여한 부분 공간에서 정상 정보를 최소화 하는 방법을 적용하였다. 일단, 티스퀘어 값을 통해 이상이 감지되면 주성분 부분 공간으로, 잔차 분석을 통해 이상이 감지되면 잔차 부분 공간에서 분석을 수행한다. 부분 공간이 정해지면, 해당 이상의 컨트리뷰션 값을 정상 상태의 주성분 분석에서 도출된 컨트리뷰션 값으로 스케일링을 수행한다. 스케일링 된 값들을 특이값 분해를 수행하여 해당하는 실시간 데이터의 센서 별 이상 크기로 새롭게 정의하여 도출할 수 있다. 도출된 센서들의 이상 크기 정보로 주요 원인 센서를 도출하고, 최종적으로 도출된 센서들을 다변수 그레인져 인과관계로 분석하여 시간이 고려된 센서 사이의 인과관계 표를 파악한다.

개발된 방법론의 성능 평가를 위해 천연 가스 액화 플랜트의 분리공정 동적 모델로 생성한 45경우의 이상 상황에 적용하였다. 제안된 이상 감시 방법론은 전체 데이터 주성분 분석 감시보다 모든 이상 상황에 대해서 월등히 빠른 감시 성능을 보였다. 단변수의 슈하트 3 시그마와 비교해서는 43개의 이상 상황에서 월등히 빠른 감시 성능을 보였다. 본 결과는 제시한 방법론이 전체 정상 데이터를 매칭되는 로칼 정상들로 정밀하게 감시함을 입증한다. 다변수 그레인져 인과관계 분석 결과를 통해서 얻은 이상 원인 정보는 전통적으로 주성분 분석의 컨트리뷰션 차트 정보

를 이용하는 것과 기존 논문의 잔차 정보를 통한 이상정보강화와 비교하였다, 그 결과로 기존 방법이 정확히 제공하지 못한 이상 원인 정보를 본 논문에서 개발한 방법론은 정확하게 제시하여 주는 것을 확인하였으며, 공정도에 결과를 도시하여 시각화된 비교를 통해 그 성능이 우수함을 확인 하였다. 본 방법론은 정상 데이터만을 기반으로 하였고, 작은 이상 상황에 맞도록 가정하여 개발하였기에 대부분의 공정에 적용 가능하며, 새로운 이상을 실시간으로 빠르게 분석하는데 크게 공헌할 것이라고 기대하는 바이다.

주요어 : 공정 모니터링, 이상 감시 및 진단, 다중 모드 운전, 그레인져 인과관계, 주성분 분석

학번 : 2011-30989

성명 : 편 하 형

감사의 글

다사다난했던, 그리고 정들었던 저의 관악 생활을 마무리 하게 되었습니다. 저의 학업 기간 동안 감사했던 분들께 이 지면을 빌어 감사의 글을 남기고자 합니다.

먼저, 대학원 기간 동안 저를 지도, 성장 시켜 주신 한종훈 교수님께 깊은 감사의 인사를 드립니다. 다양한 연구 및 사회 경험을 할 수 있도록 환경을 제공해 주시고 큰 가르침을 주셨기에 제가 졸업하여 사회로 나아갈 수 있게 되었습니다.

저의 졸업 마무리를 위해, 많이 이해해 주시고 포용해 주신 이원보 교수님께도 매우 감사 드립니다. 군복무를 외부에서 마치고, 가장 힘든 시기에 복학한 저를 응원해 주시고 지도해 주셨기에 제가 졸업할 수 있게 되었습니다. 항상 발전하는 모습을 보여 드리겠습니다.

제 학위 논문의 심사 위원장을 맡아 주시며, 논문의 완성도를 높여 주신 이규태 교수님께도 깊은 감사를 드립니다. 박사 졸업 심사에서 조언해 주셨던 조언은 회사 및 연구 생활에서 항상 실천하도록 하겠습니다.

부위원장을 맡아 주시며 연구 기본적인 부분에 대해 다시 한번 생각하게 일깨워 주신 남재욱 교수님께도 깊은 감사를 드립니다. 부족했던 부분을 늘 보완하여 발전하는 연구 활동을 하겠습니다.

부위원장을 맡아 주신 이철진 교수님께도 글로 다 표현할 수 없지만,

짧게 나마 이렇게 감사를 드리고자 합니다. 석사 신입생 때부터 대학원 생활과 프로젝트, 연구 등 다양한 방면에서 지도해 주시고, 졸업 말미에 힘들어 하던 저를 잡아 주시고, 지도해 주셔서 SCI 논문, 국내 논문, 그리고 박사 학위 논문까지 무사히 마무리 할 수 있게 되었습니다. 앞으로도 계속 좋은 모습을 보여 드릴 것을 약속 드리며, 다시 한번 깊은 감사를 드립니다.

부위원장을 맡아 주신 이웅 박사님께도 깊은 감사를 드립니다. 대학원 생활 때 제가 갖지 못했던 회사 경험, 연구에 대한 깊은 애정을 보면서 많이 배울 수 있었습니다. 항상 박사님께 배우며 발전하는 모습 보여 드리겠습니다.

제 대학원 생활 동안 공정 및 안전, 열역학에 대한 학업, 해외 학회 참석 및 발표에 대한 큰 가르침을 주시고, 석사 학위를 수여해 주신 윤인섭 교수님, 김화용 교수님, 이종민 교수님께 깊은 감사를 드립니다.

대학원 생활에서 동고동락을 함께 했던 선배님, 동기들, 후배님들께 감사 드립니다. 먼저, 저의 대학원 생활에서 영원한 큰 형님 창현이형, 항상 친근하게 대해 주시고 술도 많이 사주시며 어울려 주신 창현이형 깊은 감사를 드립니다. 연구 참여생 때 저의 첫 프로젝트 리더로 아무것도 몰랐던 저에게 많은걸 가르쳐 주셨던, 지금은 서울대 교수님이신 영섭이형, 저의 메인 프로젝트의 사수로 메인 학업을 이끌어 주신 대연이형,

연료 전지 프로젝트 때 대학원 생활의 길을 가르쳐 주시고, 학교 내외에서 항상 용기 잃지 않게 힘을 주신 현석이형, 대학원 생활의 모든 것들을 가까이서 자기 일처럼 챙겨 주신 정신적 지주 성우형, 제가 프로젝트 리더로 연구하면서 힘들어 할 때 마다 옆에서 큰 힘이 되 주셨던 상호형, 항상 긍정적인 마음으로 힘든 대학원 생활을 밝게 만들어 주신 찬샘이형, 짧게 만났었지만 대학원 입학에 큰 도움을 주신 대현이형과 지연누나, 바로 위 선배이자 친구로서 좋은 귀감을 보여 준 기욱이, 역시 선배이면서 친구로서 대학 및 대학원 생활에서 큰 도움을 준 회인이, 모든 저의 선배님들께 깊은 감사를 드립니다. 석사만 하고 졸업하였지만, 항상 웃으며 같이 배우며 프로젝트 했던 경진이, 대학생부터 항상 옆에서 같이 동고동락하며 도와 주고, 이끌어 주고, 굳은 일들을 맡아서 고생했던 정우, 먼저 졸업해서 좋은 귀감을 주기에 너무나 자랑스럽고, 고맙다는 말을 이 지면을 빌어서 글로나마 적어 봅니다.

대학생, 대학원 후배이지만 저에게 항상 본보기가 되고 항상 친하게 지낸 재흠이, 말이 필요 없는 영혼의 술친구 영수, 보기만 해도 듬직한 익환이, 엘엔지 프로젝트를 같이 잘 이끌면서 동고동락한 대근이형, 드디어 긴 터널을 같이 뚫고 나가는 만능 도움이 정남이, 고등학교부터 후배이면서 명석함을 뽐낸 시엽이, 항상 바르게 대학원 생활의 모범을 보여준 성호, 다양한 매력을 보여 주며 생활한 건희, 연구에 자부심을 느끼

며 항상 열심히 연구하는 종걸이, 남자들 사이에서 힘든 내색 없이 분위기 잘 맞춰 주며 따라 준 서린이, 즐겁게 연구 생활을 하며 항상 밝게 웃고 많이 먹는 경수. 늘 씩씩한 대장부처럼 학교 내외에서 열심히 따라 준 진주, 남녀 모두에게 인기가 많고 모든 것을 가진 남자 용석이, 졸업 논문에서 많은 것을 도와 준 능력자 창수, 자기 관리가 철저하며 스마트하게 연구한 준모, 형들의 부름에 늘 달려 나왔던 원재, 연구 테마가 달라 이제 친해 지고 있는 영근이, 마지막 저의 연구실 생활에 모든 것을 도와 준 고마운 창준이, 졸업 논문 디펜스 모든 과정을 함께하며 도와 준 용규와 가람이, 같이 생활한 적은 짧지만 스마트하게 연구했던 경우, 앞으로 남은 힘든 과정을 잘 이겨 낼 건학이, 사수인데 여러 가지 일로 제대로 챙겨 주지 못했어 서 늘 미안한 창환이, 야구 같이 하면서 늘 싱글벙글한 동우, 야구 잘하는 철원이, 힘든 대학원 생활의 분위기 메이커를 했던 민준이와 성언이, 대학원에 복귀 했을 때 항상 옆에서 도와 줬던 호동이, 가장 힘들 때 만나게 되어 제대로 챙겨 주지 못해서 미안했던 동주와 종민이, 그리고 군복무 기간에 입학해서 저를 잘 모르는데도 살갑게 대 해준 솔지, 정용이, 지영이, 6층 마지막 대학원 생활의 굳은 일을 도맡아 하며 도와 준 제 대학원 생활의 막내 재훈이, 모든 후배님들께 더 챙겨 주고 아껴 주지 못했어 서 미안하며, 저의 졸업에 큰 힘을 보태 주셔서 너무나 고맙단 말을 드립니다.

또한, 제 대학원 생활에서 영어로 대화하며 교류했던 유학생 선후배 동료인 레한, 우머, 일리야스, 크리스, 우사마에게 깊은 감사를 드립니다.

학업으로 힘든 대학원 생활에서 함께 스포츠를 즐겼던 공정 연합 – 윤인섭 교수 랩실, 김화용 교수님 랩실, 이윤우 교수님 랩실, 이종민 교수님 랩실 – 동료 및 선후배님들께도 감사 드립니다.

제 졸업의 마지막 대학원 랩실 생활을 도와 준 김효석 방장님과 이하 랩실 구성원 후배님들께도 깊은 감사를 드립니다.

사랑하는 서울대학교 화학생물공학부 동기들, 선배님, 후배님들 모두 사랑하고 감사 드립니다.

나의 배문고등학교 친구들, 특히 김도윤, 김지헌, 신상현, 유현종, 이광호, 장영찬, 항상 옆에서 응원 해주고 지켜 봐 줘서 고맙고 사랑한다.

사랑하는 교회 친구들, 은사님, 전도사님, 목사님, 합주반 제자들, 어머님들, 항상 기도 해주셔서 너무 감사 드립니다.

가장 아끼고 사랑하는 나의 양주연! 끝까지 믿고 지켜 봐 줘서 너무 고맙고 사랑해~.

마지막으로 못난 저를 끝까지 믿고 뒷바라지 해주신 저의 사랑하는 아버지, 어머니, 누나와 매형, 그리고 친척들 모두 사랑하고 감사 드립니다.