APPLICATION OF ROBUST STATISTICS ON A CRUDE DISTILLATION UNIT

by

Sinem Nalbant Kurşun

B.S., Chemical Engineering, Middle East Technical University, 2013

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering

Boğaziçi University

2017

ACKNOWLEDGEMENT

First and foremost, I would like to thank my supervisor Assoc. Prof. Burak Alakent, for his support, patience and guidance during my graduate study. Without his valuable assistance, this work would not have been completed. I feel very lucky to have had the opportunity to work with him and learn from him.

I would also like to thank my committee members Prof. Dr. Ramazan Yıldırım and Assist. Prof. Kazım Yalçın Arga for devoting their valuable time to read, listen and comment on my thesis and supporting me to graduate.

My deep gratitude goes to my family for their love and support over the years. My mother Ayşe, my father Mustafa and my sister Seren have always encouraged me unconditionally and they always been there for me. I am thankful for everything they helped me to conclude my work.

My private thanks are for my beloved husband, Tolga, for his never ending encouragement and the faith in me. I have always felt his love and moral energy deep in my heart.

Also, I want to thank to my colleagues in TUPRAS for their cooperation, technical support and understanding during preparation of this work.

Finally, BAP Project No. 8041 is gratefully acknowledged.

ABSTRACT

APPLICATION OF ROBUST STATISTICS ON A CRUDE DISTILLATION UNIT

Refineries are highly complex and integrated systems, separating and transforming crude oil into valuable products. One of the most important processes in refineries is the Crude Distillation Unit (CDU) process, in which raw crude oil is separated into various fractions to be further processed in other parts of the refinery. In the refinery, Heavy Diesel (HD) T95 value is very important quality indicator. In the current study, conventional and robust statistical methods were employed on the historical data of a CDU process in TUPRAS İzmit Refinery for monitoring and HD T95 prediction purposes. Process data consisted of online measurements of process variables and laboratory measurements of HD T95 values for a one-year period. In the first part of the study, trajectories of process variables were analyzed to identify relations between process variables and to distinguish normal from abnormal operating conditions in the distillation history. For this purpose, skipped-Principal Components Analysis (PCA) and Minimum Covariance Determinant (MCD)+PCA methods were applied to process data and MCD+PCA method was found as more efficient method in detecting disturbances in the operation conditions. In the second part of the study, Monte Carlo (MC) simulations were applied by creating clean and contaminated datasets to evaluate predictive performances of LS and various robust regression methods, and to assess the metrics (RMSE, MAE) for evaluating the quality of predictions under contamination. LTS10%+LS and LTS20%+LS were found as best predictive models, and RMSE was found to be reliable in assessing models when 70%-90% of the highest absolute prediction errors were taken into account. In the last section, LS and robust regression methods were applied and compared to select the most convenient prediction method for HD T95 values. The best predictive performance was obtained by LTS30% model with 97.5% CL. By applying this method to historical dataset, 15% of training dataset was detected as outliers and when these outliers were excluded from dataset, the model can predict HD T95 value with a maximum 7 ^{0}C error.

ÖZET

HAM PETROL DESTİLASYON ÜNİTESİNDE SAĞLAM İSTATİSTİKLERİN UYGULANMASI

Rafineriler, ham petrolü ayrıştıran ve değerli ürünlere dönüştüren oldukça karmaşık ve entegre sistemlerdir. Rafinerilerdeki en önemli süreçlerden birisi ham petrolün rafinerinin diğer bölümlerinde işlenmek üzere ayrıldığı Ham Petrol Destilasyon Ünitesidir (CDU). Ağır Dizel (HD) T95 değeri rafineride önemli bir kalite göstergesidir. Bu çalışmada, süreç değişkenlerini izleme ve HD T95 tahmini için geleneksel ve dayanıklı istatistiksel yöntemler, TÜPRAS İzmit Rafinerisi CDU süreci geçmiş verileri üzerinde uygulanmıştır. Süreç verileri, bir yıllık döneme ait 23 süreç değişkeninin çevrimiçi ölçüm değerlerini ve HD T95'in laboratuvar ölçüm değerlerini içermektedir. Çalışmanın ilk bölümünde, süreç değişkenleri arasındaki ilişkileri tespit edebilmek ve destilasyon sürecindeki anormal çalışma koşullarını ayırt edebilmek için süreç değişkenleri arasındaki ilişkiler analiz edilmiştir. Bu amaçla, süreç verisine Atlanan-Temel Bileşenler Analizi (PCA) ve En Küçük Varyans-Kovaryans Determinantı (MCD)+PCA yöntemleri uygulanmıştır. MCD+PCA yönteminin çalışma koşullarındaki bozuklukların tespitinde daha etkili olduğu tespit edilmiştir. Çalışmanın ikinci bölümünde, En Küçük Kareler yöntemi (LS) ve çeşitli dayanıklı regresyon yöntemlerinin tahmin edici performansları ve kontaminasyon altında tahminlerin kalitelerini belirlemek için RMSE ve MAE metriklerinin kullanım uygunluğunu değerlendirmek amacıyla temiz ve kontamine veriler oluşturularak Monte Carlo (MC) benzetimleri yapılmıştır. En iyi tahmin edici modeller En Küçük Kırpılmış Kareler (LTS) 10%+ En Küçük Kareler (LS) ve LTS20%+LS olarak bulunmuştur. Ayrıca, en yüksek mutlak tahmin hatalarının %70-90'ı dikkate alındığında, RMSE'nin daha güvenilir bir değerlendirme yöntemi olduğu belirlenmiştir. Son bölümde, HD T95 değerleri tahmininde en uygun tahmin yöntemini belirlemek için LS ve dayanıklı regresyon yöntemleri uygulanmış ve karşılaştırılmıştır. En iyi tahmin performansı, %97,5 güven düzeyinde %30 kırpma ile (LTS) yöntemiyle elde edilmiştir. Bu yöntemin tarihsel veri seti üzerine uygulanmasıyla, veri setinin %15'i aykırı gözlem olarak tespit edilmiş ve bu gözlemler veri setinden çıkarıldığı zaman, HD T95 değeri en çok 7 ^{0}C hata ile tahmin edilebilmektedir.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

ANN	Artificial Neural Networks
BDP	Breakdown Point
CCR	Continuous Catalytic Reactor
CDU	Crude Distillation Unit
CL	Confidence Limit
CV	Cross Validation
DHP	Diesel/Kerosene Hydroprocessing
HD	Heavy Diesel
HSRN	Heavy Straight Run Naphta
HVGO	Heavy Vacuum Gas Oil
LD	Light Diesel
LMS	Least Median Squares
LPG	Liquified Petroleum Gas
LR	Linear Regression
LS	Least Squares
LSRN	Light Straight Run Naphta
LTA	Least Trimmed Absolute Value
LTS	Least Trimmed Squares
LVGO	Light Vacuum Gas Oil
MAE	Mean Absolute Error
MC	Monte Carlo

MCD	Minimum Covariance Determinant
MD	Mahalanobis Distance
MLR	Multiple Linear Regression
MVA	Multivariate Analysis
MVE	Minimum Volume Ellipsoid
OLS	Ordinary Least Squares
PC	Principle Component
PCA	Principle Component Analysis
PCR	Principle Component Regression
PLS	Partial Least Squares
PRESS	Prediction Residual Sum of Squares
RMSE	Root mean Squared Error
SE	Standard Error of Coefficient
SPC	Statistical Process Control
VDU	Vacuum Distillation Unit
WLS	Weighted Least Squares

1. INTRODUCTION

A crude oil is a naturally occurring, unrefined petroleum product, composed of hydrocarbon deposits. It is trapped in different reservoirs of the world, and according to the region of reservoir, chemical and physical characteristics of crude oil changes [1]. Classification of crude oil is based on the differences in specific gravities and the proportions with which it is formed. Product demand is met by blending different crudes as required proportion [2].

Petroleum refineries are highly complex and integrated systems, separating and transforming crude oil into a wide variety of high value products with respect to boiling range and carbon distribution [3]. The main products are liquefied petroleum gas (LPG), gasoline, kerosene, diesel and fuel oil. In many refineries, crude oil is not only distilled but also converted and blended into different products. Refineries consist of different process units, such as Crude Distillation Unit (CDU), Continuous Catalytic Reactor Unit (CCR), Isomerization Unit, Diesel/Kerosene Hydroprocessing Unit (DHP), Hydrodesulfurizer Unit, LPG-Amine Treating Unit.

The first process unit of the petroleum refinery is CDU, in which raw crude oil is separated, based on their boiling points, into various fractions, each of which is then processed further in other parts of the refinery. Boiling point is a reliable indicator of the molecular weight (or length of the carbon chain) of different products. In CDU, the top, or lightest fractions comprise fuel gas, LPG and gasoline. The middle fraction is made up of kerosene and diesel, while and the heaviest fraction mainly consists of fuel oil.

Process control has a significant role for safe and efficient operation of CDU. Using improved process control systems, a refinery plant can be operated closer to optimum values [5]. To collect data for process control, refinery is instrumented with a large number of on-line sensors, measuring temperatures, pressures, and flow rates, as frequent as every second. The resulting data from these measurements are high-dimensional and possibly time-dependent, and stored in a database, which may be used as a reference dataset for further studies. Multivariate process monitoring methods are used for dimensional

reduction and analysis of the historical data. Using multivariate statistical methods, on-line monitoring of the process may be performed to identify process failures, and improve process quality [19].

In refinery process units, process variables, such as temperature, pressure, flow rates, and quality variables, such as gravity, flash point, boiling points, viscosity, are monitored to evaluate the performance of the processes. Process variables directly related with product quality cannot usually be measured on-line by hardware sensors, since on-line measurements of quality indicators are expensive, slow and sometimes unreliable. In this case, quality variables are measured at a lower and variable frequency, compared to the sampling of process variables [7], usually determined by laboratory analyses, which are performed periodically such as 1-3 times in a day, or sometimes 1-3 times in a week. Thus, it is not a straightforward task to monitor the final product quality on-line, and control the quality variable [8].

Overall structure of the refinery industry has changed in recent years because of a growing demand for lighter products; demand for heating (fuel) oil is decreasing, and demand for gasoline, jet fuel and diesel is increasing [3]. This demand has led to more complex refineries with increased conversion capacities. Increased conversion will unavoidably lead to increase in energy consumption, but will also yield a product with a higher quality. Refinery products must be synchronous with sales specifications based on product qualities, in order to be sold. When the market demand and process requirements in the refinery are taken into consideration, operational efficiency and economic benefit highly depend on accurate prediction of the product qualities. Since on-line measurement of quality variables, as discussed above, is not an easy task, on-line measured process variables are used for on-line prediction of the product quality variables. Functional relation between the process and quality variables is estimated, and the resulting “soft-sensor” model may be used to predict the level of a quality variable, -corresponding to a set of measured process variables, without the need for laboratory analysis. Using soft sensors, operators can assess the process performance on-line, and they can adjust the manipulated variables in the process in the correct direction to obtain the desired level of the quality variable. Furthermore, soft-sensors may also be used by automatic controllers.

A significant difficulty in measuring quality variables lies in low reliability of the laboratory measurements compared to on-line process measurements. Although it is a common assumption, held by many engineers and researchers, that measurement errors are normally distributed, there may be "contaminated" data, particularly in the laboratory measurements. Here, contamination in data refers to existence of outliers, which come from an error distribution different than the error distribution of the "clean" data. Contaminated data may result from inhomogeneity in experimental equipment, personnel and conditions, and may be basically revealed as biased measurements, or heteroscedastic variance. Application of least-squares (LS) on linear regression (LR) models is a common technique for constructing soft-sensor models from historical data, and a convenient method when measurement errors are normally distributed. In the presence of outliers, however, regression model parameters determined by LS may be biased and/or have high variance, hence predictions from LS models may be unreliable and misdirect the operators and the control system. When data are high dimensional, it is not straightforward to identify outliers, so alternatives to LS regression is required. The main focus of the current thesis is to construct soft-sensor models from real industrial data sets, possibly contaminated with outliers, for process monitoring and quality prediction, using robust statistical techniques.

The current study involves CDU in TUPRAS İzmit Refinery, in which fuel gas, LPG, naphtha, kerosene, diesel and fuel oil are fractionated according their boiling points. These products are sent to the storage, or to other process units as feedstock. Diesel product is fractionated as light diesel (LD) and heavy diesel (HD). HD T95 value, which is the temperature at which 95% of diesel volume is evaporated, is an important variable for HD. T95 value is measured once in a day by laboratory analysis. Although quality tests are repeated for at least three times in the TUPRAS laboratory for products offered for sale, HD T95 measurements is performed once, since HD in CDU is an intermediate product to be processed later. Therefore, accurate on-line predictions of HD T95 is crucial for the operation to be adjusted for the sales specifications.

In the current thesis, conventional and robust statistical tools are performed on the historical data of CDU process in TUPRAS İzmit Refinery for process monitoring and quality (HD T95) prediction. Sections 2 to 4 consist of three descriptive sections: i)

description of conventional and robust linear statistical methods used to construct soft-sensor models from historical data, ii) process description, in which refinery, distillation process and the CDU of TUPRAS İzmit Refinery are discussed, and iii) an overall description of the historical dataset. Section 5 is the Results and Discussion, which consists of three subsections: i) process monitoring using historical data, ii) Monte Carlo (MC) simulations to asses various models and metrics in the presence of contaminated data, and iii) predictive model construction using historical data. Main findings and suggestions for further studies are summarized in the Conclusion section.

2. LINEAR STATISTICAL METHODS FOR EXPLORATORY AND PREDICTIVE MODELING

Modeling is a representation of a system, or a process, aimed to give some information on how something behaves in real life. There are mainly two types of modeling: mechanistic and statistical modeling. Mechanistic models are the models based upon fundamental principles. For the complex processes, constructing mechanistic models is hard and time-consuming to develop. Statistical models are data based models, using statistical relationship between the variables in the historical data set. They are easy to develop and more practical than mechanistic models. Statistical model building, based on historical plant data, is usually the most cost-effective way to obtain a process model. There are various statistical modeling and analysis methods. The most popular techniques are the Multiple Linear Regression (MLR), Principle Component Analysis (PCA), Partial Least Squares (PLS), Artificial Neural Networks (ANN), Neuro-Fuzzy Systems, Support Vector Machines. In this section, only linear statistical methods will be discussed. Furthermore, robust exploratory and predictive modeling techniques such as Least Median Squares (LMS), Least Trimmed Squares (LTS), Minimum Covariance Determinant (MCD), M-Estimator and S-Estimator will be discussed [11].

2.1. Least Squares (LS) Analysis

LR is a statistical method used to model the relation between dependent and independent variables. As reported by a study on Japanese chemical and refining industries, LR models yield sufficient prediction accuracy for most distillation and reaction processes [17]. LR may be classified as simple LR and multiple LR, and LR model parameters are conventionally determined using the LS method. LS is one of the oldest and easiest techniques of the modern statistics. LS analysis is used to estimate regression parameters by minimizing the squared error between the observed data and fitted values. LS is developed in the late 1700's and the early 1800's by a number of mathematicians; Adrien Marie Legendre, Karl Friedrich Gauss and (possibly) Robert Adrain working in France, Germany and America, respectively [12]. LS was firstly published by the French mathematician Adrien Marie in 1805 in his work “Nouvelles Methodes pour la

Determination des Orbites des Cometes" [13]. The technique is described as an algebraic procedure for fitting linear equations to data. In 1809, the German mathematician Karl Friedrich Gauss published his method of calculating the orbits of celestial bodies, in which he claimed that he has previously discovered LS and used it as early as 1795 in estimating the orbit of an asteroid [14]. In 1886, Galton used LS in his work on the heritability of size, laying down the foundations of correlation and regression analysis [14]. Pearson and Fisher, who did so much in the early development of statistics, used and developed it in different contexts (factor analysis for Pearson, and experimental design for Fisher) [14].

LS analysis is widely used to estimate the parameters of a function fit to a dataset, and to characterize the statistical properties of estimates [14]. By using LS, the best fitting line can be found by minimizing the sum of the squares of the vertical distance from each data point on the line [15]. The simplest version of the LS is Ordinary LS (OLS) and more sophisticated version is Weighted LS (WLS).

OLS is a method for estimating the unknown parameters in a LR model by minimizing the sum of the squares of the error, which is the vertical distance between data points and the regression line. OLS method is used to minimize error E by determining estimators β_0 and β_1:

$$E = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i [Y_i - (\beta_0 + \beta_1 X_i)]^2 \tag{2.1}$$

Here $\{Y_i, X_i\}$ is a set of N pairs of observations where Y represents dependent (response) variables, X represent independent (predictor) variables. OLS estimators are linear, unbiased and have small variances. However, when error does not have a constant variance, WLS method, also called as Generalized LS, provides better estimate [14].

In WLS, the idea is to assign each observation a weight, w_i that reflects the uncertainty of the measurement. Weight w_i is a function of the variance of the i^{th} observation. In many cases, the variance σ_i^2 depends on x_i. Observations where σ_i^2 is large are less accurate so they should play a smaller role in the estimation of β_1 [16]. As in OLS method, WLS method also aims to minimize error E_w to find estimators.

$$E_w = \sum_i w_i (Y_i - \hat{Y}_i)^2 = \sum_i w_i [Y_i - (\beta_0 + \beta_1 X_i)]^2 \quad (2.2)$$

2.1.1. Simple Linear Regression

The simple LR model is used to find the straight line that best fits the data. It illustrates the relation between the dependent variable *y* and the independent variable *x* based on the regression equation. In simple LR, the error distribution is assumed to be normal. A simple LR model can be written as:

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (2.3)$$

where X and ϵ, usually taken to be measurement error, represent the independent variable and the random error, respectively, and ϵ has mean 0 and standard deviation σ.

In fitting a line to a data set using simple LR, it is aimed to find coefficients β_0 and β_1 minimizing the square of the errors between observed and the fitted values of *Y*. The fitted value of *Y* for a given *X* is determined as follows:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X \quad (2.4)$$

Here, $\hat{Y}$ is predicted response variable, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the estimates of the regression parameters.

2.1.2. Multiple Linear Regression (MLR)

MLR is used to fit a model with two or more independent variables. In MLR analysis, the method of LS may be used to estimate the regression coefficients. The regression coefficients represent the partial contribution of each predictor variable to the response variable. Unlike the simple LR, contribution of the order of independent variables and interactions between independent variables should be taken into consideration. In fitting MLR, the aim is to find the best estimates of the coefficients via minimizing the

sum of squares of the errors between observed and the fitted responses. The general equation of MLR model for *k* variables is [15]:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i, \quad i = 1,2, \ldots \ldots, n \tag{2.5}$$

The MLR model can also be written in matrix notation as:

$$Y = X\beta + \epsilon \tag{2.6}$$

where Y is a $n\times1$ dimensional random vector, X is an $n\times(k+1)$ matrix, consisting of predictors, β is a $(k+1)\times1$ vector of unknown parameters, and ϵ is an $n\times1$ vector of random errors. To be able to determine the LS estimates of the model coefficients ($\hat{\beta}$), Equation 2.7 should be solved:

$$X^T X\hat{\beta} = X^T Y \tag{2.7}$$

By multiplying both sides with $(X^{\mathrm{T}}X)^{-1}$, we obtain [15]:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.8}$$

Mahalanobis distance (MD) is used to determine the scaled distance of a sample point to the centroid of all the data points in the predictor space. MD depends on the variance of each variable and covariance between process variables. MD of each measurement x_i is calculated by following formula.

$$MD(x_i) = \sqrt{(x_i - \hat{\mu}_0)' \hat{\Sigma}_0^{-1} (x_i - \hat{\mu}_0)} \tag{2.9}$$

where, $\hat{\mu}_0$ is the arithmetic mean and $\hat{\Sigma}_0$ is classical covariance matrix [18]. Data points with high MD statistics are likely to affect the LS model more.

Data points making substantial differences in regression models are called influential points. Influential points are outliers and leverage points. Outliers are the observations that are far from the pattern described by other data points. In industry, these data points may be produced as a result of operation conditions, measurement errors, etc. LS method is highly sensitive to outliers, and even one outlier may change the fitted plane dramatically. It is difficult to detect outliers with the exploratory data analysis techniques. Leverage points are observations deviating from the majority in x-space, i.e. high MD. Leverage points are categorized according to their place on regression line. “Good” leverage points follow the pattern of the majority data points, while “bad” leverage points do not follow this pattern [18]. For the low dimensional datasets, outliers and leverage points can be detected by visual inspection. However, this is not a straightforward task for high dimensional datasets. As discussed below, robust estimators are needed for the reliable analysis of the regression in the presence of outliers.

2.2. Principle Component Analysis (PCA)

Statistical process control (SPC) is a tool for detecting changes in industrial processes, achieving and maintaining product quality [7]. The most widely used SPC techniques are univariate analysis and multivariate analysis (MVA) methods. Univariate analysis is the simplest form of statistical analysis; individual quality measurements are collected, these measurements are visualized and analyzed sequentially [7]. MVA is the analysis of more than one variable, and comprised of exploratory data analysis, classification, regression analysis and predictive modeling methods. In MVA, all data are analyzed simultaneously, and directionality of the common variations in the variables is used to extract information from the process.

PCA is probably the oldest and best-known technique in multivariate analysis. This technique was first introduced by Karl Pearson 1901, and later developed by Harold Hotelling in 1930s. PCA is generally accepted as revolution in the use of multivariate methods [20]. Nowadays, it is generally used in statistical data analysis, communication theory, pattern recognition, and image processing. In chemical processes, measurement of process variables, such as temperature, flow, pressure by sensors yield high-dimensional and time dependent data, which may be challenging for SPC. PCA is a mathematical

transformation method, used to identify the patterns, and reduce the dimensions in plant measurements data [21].

PCA is utilized to extract the components with the largest variances, to explain and simplify the description of the dataset, to analyze the structure of observations and variables, to detect outlier observations and for warning for potential malfunctions [21]. PCA is used to compute new variables called principal components (PCs). The uncorrelated PCs are weighted sum of the original process variables. The first PC is the linear combination of the process variables having largest variance to describe greatest amount of variability of dataset. The second PC is the linear combination of the process variables having next largest variance, and subject to the constraint of being orthogonal to the first PC [7]. This procedure is repeated until an adequate number of PCs is obtained in the same way.

In PCA application, the mean-centered data matrix X is of size (n x m), in which n is the number of samples, and m is number of process variables. Using eigenvalue decomposition on $X^T X$, l eigenvectors (P), which correspond to the largest l eigenvalues, are selected. Hence, data matrix X may be decomposed into the sum of the outer product of l pairs of vectors:

$$X = TP^T + E \tag{2.10}$$

where, T (n x l), P (m x l) and E (n x m) are PC scores, loadings and residuals matrix, respectively, while l is the number of PCs [22]. A data vector of original dimensions x_i (1 x m) may be reconstructed in the PC subspace, using the following transformation formula:

$$\hat{x}_i = t_i P_l^T = x_i P_l P_l^T \tag{2.11}$$

where, t_i is a vector of scores for sample x_i [22]. Illustration of the PCA method is given in Figure 2.1.

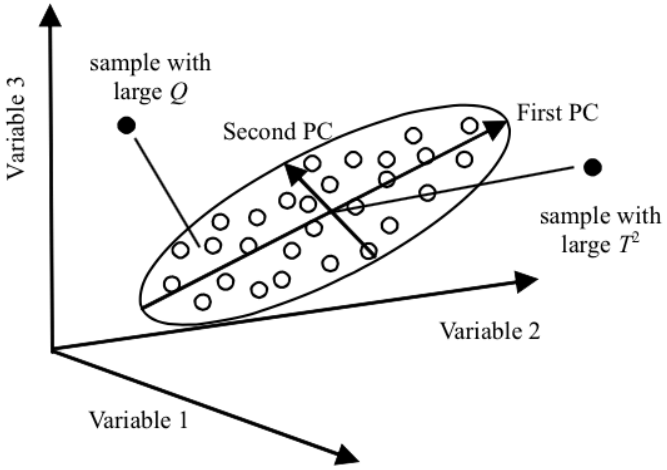

Figure 2.1. Data projection on two PCs [22].

There are mainly two statistics used to evaluate PCA models: Hotteling's Q and T^2 statistics. Q statistics shows how well a single observation is fitted by PC plane. It is calculated by taking the squared difference between the original data and its projection on PC plane. T^2 statistic is used to measure the variation within the PC subspace. T^2 statistic is calculated by the summation of the squares of the adjusted (unit variance) scores on each of the PCs in the model [22]. Observations falling outside the confidence limits of Q and T^2 statistics are deemed to be outliers, and generally removed from the dataset.

$$Q_i = r_i r_i^T \tag{2.12}$$

Here, $r_i = x_i - \hat{x}_i$.

$$T^2 = \sum_{i=1}^{l} (\frac{t_i}{\lambda_i})^2 \tag{2.13}$$

PCA is known to be sensitive to outliers. It should be noted that outlier detection potential of PCA is deteriorated by the very same outliers in the dataset, i.e. outliers in the reference set perturb the estimation of the true model hence, outliers cannot be detected by the constructed model. Robust PCA procedures have been developed to overcome this problem [23].

2.3. Multivariate Robust Statistics and Robust Regression Methods

Robust regression is a statistical method designed to increase stability of estimates and reliability of models in the case that parametric model assumptions are not valid. Robust regression methods have been developed to improve the performance of LS, in the presence of outliers [24]. Robust regression estimators are not easily affected by outliers, hence provide stable fits even in the presence of outliers. In order to achieve this stability, robust regression, basically, limits the influence of outliers (Figure 2.2).

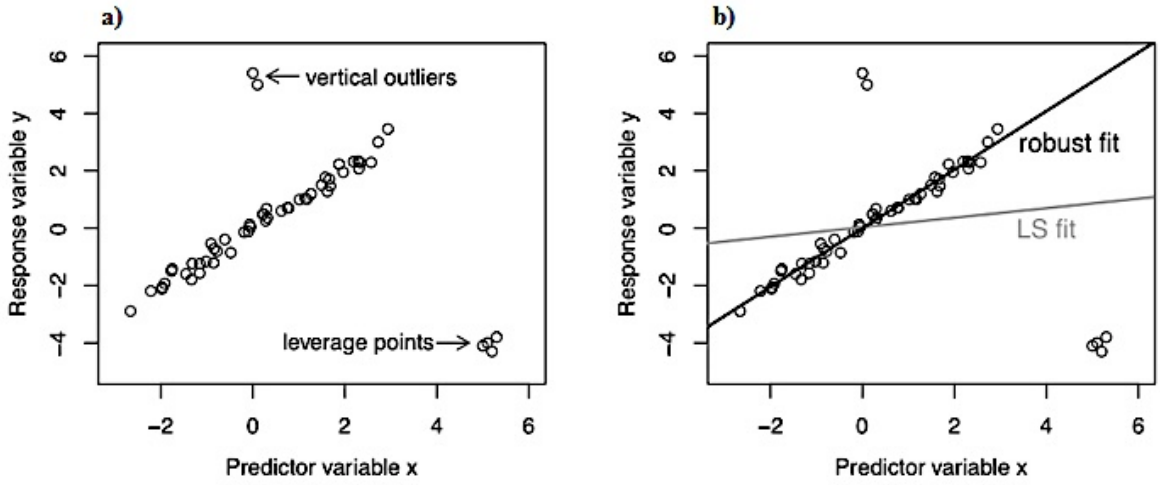

Figure 2.2. (a) Linear regression problem with vertical outliers and leverage points, and (b) result of LS regression and robust regression [25].

Three important properties of robust estimators, breakdown point, influence function and efficiency, are discussed below.

Breakdown point (BDP): The breakdown point is defined as the smallest fraction of anomalous data that can render estimator useless [26]. The BDP deals with the problem of large contamination. It characterizes the smallest amount of contamination that can cause an estimator to yield arbitrary values. Robust estimators have a positive BDP, meaning that a certain part of the data could be "outliers", and the estimator gives still useful results [25]. For sample $X = (x_i, y_i)$, BDP $\epsilon^*(T, X)$ of the estimator T at X is calculated as [27]:

$$\epsilon^*(T, X) = min\{m/n\,;\beta(m; T, X) is\ infinite\} \tag{2.14}$$

Here, *m* and *n* refer to the number of contaminated and total observations, respectively. High-BDP regression estimators have been developed to provide reliable estimates in the presence of a large percentage of outlying observations. These estimators can achieve up to a 50% BDP and are also known as resistant estimators. They are useful for outlier detection and initial estimators [26]. Different regression methods have different BDPs, e.g. LS has, a BDP of 0, while BDP is 50% for Least Median of Squares (LMS).

Influence function: The influence function is used to examine the response of an estimator upon an infinitesimal contamination [25]. Robust estimators ideally have a bounded influence function, which means that there is a small effect on the estimator.

Statistical efficiency: The efficiency is defined as the ratio of the variances of OLS estimator and the robust estimator in the absence of outliers, usually for normal distributed data. It can be shown that the efficiency is in the interval 0-1, where 1 refers to highly efficient estimator [25].

There are various types of robust regression estimators: Theil–Sen estimator, Least Median of Squares Estimator (LMS), Least Trimmed Squares Estimator (LTS), Least Trimmed Absolute Value Estimator (LTA), M-Estimators, MM-Estimators, S-Estimators etc. [28]. Most of these estimators work on a similar principle; a smaller weight is given to observations that would otherwise influence the regression line.

2.3.1. Least Median Squares (LMS)

LMS originate to Tukey who proposed an estimator based on the shortest half of the sample [29]. Then, Hampel modified and generalized it to regression and stated that the resulting estimator has a 50% BDP [30]. However, LMS was firstly proposed, in its full form, by Rousseeuw (1984). Rousseeuw provided the theory and algorithm for estimators having 50% BDP. LMS was further developed by Rousseeuw and Leroy (1987) [31]. While the mean of the squared residuals is minimized in LS regression, mean of the squared residuals is replaced with the median of the squared residuals in LMS regression. Hence, the median-based LMS parameter estimators are resistant to outliers.

For the dataset including n observations and p parameters, $(x_i, y_i) = (x_{i1}, \dots, x_{ip}, y_i)$ is (p+1) dimensional linear space. The estimates β of LMS are found by solving the optimization equation:

$$\min med\,(y_i - \sum x_{ip}\beta_p)^2 = \min med(\epsilon_i^2) \tag{2.15}$$

where ϵ_i denotes the i^{th} residual [32]. If the number of parameters is higher than 1 (p>1), and the observations are in general position, BDP of LMS method is calculated with the following formula [27]:

$$([n/2] - p + 2)/n \tag{2.16}$$

Here [.] represents the integer part of the quantity inside the squared brackets. Although LMS has high BDP, its asymptotic efficiency relative to the OLS estimator is 0. Also, it does not have a well-defined influence function because of its low convergence rate $n^{1/3}$ [32]. Hence, especially for high-dimensional datasets and high number of observations, computation time is too long. Due to these disadvantages of LMS method, Least Trimmed Squares (LTS) method may be used.

2.3.2. Least Trimmed Squares (LTS)

LTS method was introduced by Rousseeuw in 1983 [33]. The objective function of LTS method is defined as:

$$\hat{\beta}_{LTS,h,N} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{h} r_{(i)}^2(\beta) \tag{2.17}$$

where $r_i(\beta) = y_i - \beta^T x_i$ [34]. For a dataset with n observations, trimming parameter h is selected, such that $([(n + p + 1)/2] \leq h \leq n)$. Generally, trimming parameter h can be based on the trimming proportion α:

$$h = [n(1-\alpha)] + 1. \tag{2.18}$$

In LTS method, residuals are written in ascending order: $|r_{(1)}| \leq |r_{(2)}| \leq \cdots \leq |r_{(n)}|$, and the LTS estimator is determined via minimizing the sum (or mean) of the smallest *h* squared residuals. For small datasets, exact solution to the LTS estimator is easily determined, but this is not a simple task for large datasets. Rousseeuw and Leroy determine LTS estimators, using exact fits on subsets of size *k*, representing the number of total parameters in the model [33]: When the number of the observations is small enough, the following procedure can be applied for computation of LTS [34].

- Firstly, the trimming parameter, *h* is selected.
- All possible subsets with *k* observations are generated, and regression parameters are computed via solving each exact fit equation.
- By using all observations, residuals are calculated, and LTS criterion is applied on the ranked square residuals.
- The LTS objective function to estimate of LTS ($\hat{\beta}_{LTS,h,N}$) is applied.

Taking $h = [(n + p + 1)/2]$, LTS regression has ~50% BDP. When $h = n$, LTS regression is equivalent to the standard LS regression. LTS estimator converges at a rate of $n^{1/2}$, faster than that of LMS estimator. LTS is usually preferred over LMS, since LTS converges faster and has smoother objective function [34]. However, LTS suffers from low efficiency and computation complexity. Due to its low efficiency, LTS may be suggested as starting point for more efficient procedures or estimators.

Due to low efficiency of LTS method, some additional procedures are proposed to increase efficiency. The following method suggested by Rousseeuw and Hubert is called "reweighted LTS" in the current study [34]:

- Firstly, trimming parameter *h* is selected to obtain estimate $\hat{\beta}_{LTS,h,n}$ and the variance is derived from LTS.

$$\hat{\sigma}_0^2 = \frac{1 - \alpha}{1 - \alpha - 2c_\alpha \phi(c_\alpha)} \hat{\sigma}_{LTS,h,n}^2 \tag{2.19}$$

where $\alpha = (n - h)/n$ and $c_\alpha = F^{-1}(1 - \frac{\alpha}{2})$.

- After constructing estimated variance; t-statistics for all *n* residuals are constructed:

$$u_{0,i} = (y_i - x_i' \hat{\beta}_{LTS})/\hat{\sigma}_0 \tag{2.20}$$

- To find an updated scale estimate, use the following weights:

$$\varpi_i = I\left[\left|u_{0,i}\right| \leq c_2\right] \qquad i = 1, \ldots, n \tag{2.21}$$

where *I*[.] is the indicator function, and c_2 is usually taken to be equal to 2.5. In the current study, c_2 is taken to be equal to $F^{-1}(1 - \frac{0.025}{2})$.

$$\hat{\sigma}_1^2 = \frac{1}{\sum_{i=1}^n \varpi_i - k} \sum_{i=1}^n \varpi_i \, (y_i - x_i' \hat{\beta}_{LTS})^2 \tag{2.22}$$

- then, new t-statistics are:

$$u_{1,i} = (y_i - x_i' \hat{\beta}_{LTS})/\hat{\sigma}_{1.} \tag{2.23}$$

- By selecting observations according to $\left|u_{1,i}\right| \leq c_2$ criteria, a clean sample is obtained and OLS method is applied this data [34].

2.3.3. M and GM-Estimators

Being one of the first robust regression methods, M-Estimator was proposed by Huber in 1964 [18]. “M“ in M-Estimator indicates maximum likehood estimation. M-Estimator is an alternative robust estimator to the LS. It can also be considered as a WLS

method, in which weights are determined by the residuals [35]. In this method, e^2 in the LS objective function is replaced with a symmetric, positive-definite and minimum at zero weight function of residuals, $\rho(e)$. The aim of the M-estimation is to minimize the sum of the residual function:

$$Min \sum_{i=1}^{n} \rho(e_i). \tag{2.24}$$

In Generalized M Estimators (GM-Estimators), model residuals are scaled with respect to their leverage values, named Schweppe weighting:

$$r = e/(k \times s \times \sqrt{1-h}) \tag{2.25}$$

where h is the vector of leverage values, and s is an estimate of the standard deviation of the error term [36].

M-estimators can also be derived in the form of WLS. Weighting function ρ can be taken in various forms. Table 2.1 shows weighting functions commonly used in MATLAB.

Table 2.1. Weighting functions in MATLAB [36].

Weight Function	**Equation**	**Default k**
'andrews'	$w = (abs(r) < \pi) \times \sin(r)/r$	1.339
'bisquare'	$w = (abs(r) < 1) \times (1-r^2)^2$	4.685
'cauchy'	$w = 1/(1+r^2)$	2.385
'huber'	$w = 1/\max(1, abs(r))$	1.345
'logistic'	$w = \tanh(r)/r$	1.205
'welsh'	$w = \exp(-(r^2))$	2.985

2.3.4. S-Estimator

S-Estimators of regression was proposed by Rousseeuw and Yohai (1984) [37]. S estimator is the generalization of LMS and finds a line minimizing a robust estimate of the scale of the residuals [38]. It is called as "S ", because it is based on estimators of scale. The method is highly resistant to outliers, yielding a high BDP, and S-estimator has a convergence rate of $n^{-1/2}$. S-estimators show same asymptotic properties as M-estimators [38]. The S-regression parameter estimates minimize the following objective function:

$$Min\ s(\theta) \tag{2.26}$$

where $s(\theta)$ is certain type of robust M-estimate of the scale of the residuals $e_i(\theta), \ldots, e_n(\theta)$, and satisfying the following constraint:

$$K = \frac{1}{n}\sum_{i=1}^{n} \rho(\frac{e_i}{s}) \tag{2.27}$$

where, *K* is taken to be $E_\phi[\rho]$, and ϕ is the standard normal distribution [37]. For a specific BDP, *K* value is determined, Equations 2.26 and 2.27are solved to yield S-regression parameter estimates.

2.3.5. Minimum Covariance Determinant (MCD)

It is usually difficult to determine location and scatter of high-dimensional datasets using robust estimators. In 1984, Rousseeuw has introduced the MCD estimator, which is a highly robust estimator of multivariate location and scatter [18]. MCD estimator is used to determine *h* observations out of *n* observations, with the classical covariance matrix having the smallest determinant. The MCD estimate of location (T) is the mean of the determined *h* observations, and the MCD estimate of the scatter (S) is the covariance matrix of *h* observations [39].

An iterative procedure is used to compute MCD estimator. Taking a dataset $X_n = \{x_1, \dots, x_n\}$ with p variables, and is defining $H_1 \subset \{1, \dots, n\}$ with $H_1 = h$ The location and scatter estimates are computed as follows [39].

$$T_1 = \frac{1}{h} \sum_{i \in H_1} x_i \tag{2.28}$$

$$S_1 = \frac{1}{h} \sum_{i \in H_1} (x_i - T_1)(x_i - T_1)' \tag{2.29}$$

If $\det(S_1) \neq 0$, then the the relative distances are defined:

$$d_1(i) = \sqrt{(x_i - T_1)' S_1^{-1} (x_i - T_1)} \quad \text{for } i = 1, \dots, n. \tag{2.30}$$

Next, $H_2$is taken, with the condition that $\{d_1(i);\ i \in H_2\} = \{(d_1)_{1:n}, \dots, (d_1)_{h:n}\}$, where $(d_1)_{1:n} \leq (d_1)_{2:n} \leq \cdots \leq (d_1)_{n:n}$ are the ordered distances. Using the data points with the indices in H_2, T_2 and S_2 are computed. Then, the following condition is checked,

$$\det(S_2) \leq \det(S_1) \tag{2.31}$$

The equality is obtained if and only if $T_2 = T_1$ and $S_2 = S_1$. Repeating an iterative process, the algorithm is stopped if $\det(S_1) = 0$, or $\det(S_1) = \det(S_2)$.

As mentioned in Section 2.2, PCA is probably the oldest and best-known technique of multivariate analysis. However, PCA method is sensitive to outliers so it may produce unreliable results if dataset contains outlier observations [23]. Hence, to eliminate the effect of outliers in PCA application, robustification methods are used. MCD is a well-known PCA robustification method in the literature.

3. PROCESS DESCRIPTION

In this section, refinery sector (Section 3.1), distillation theory (Section 3.2), CDU in TUPRAS İzmit Refinery (Section 3.3) and laboratory test method of HD T95 (Section 3.4) are briefly discussed.

3.1. Refinery Sector

Petroleum is a naturally occurring complex mixture of organic liquids, consisting of crude oil and natural gas. Petroleum is composed of different organic hydrocarbon molecules, such as paraffins, naphthenes, aromatics, asphaltenes, and also contains nitrogen, oxygen, sulfur and some metals like iron, nickel and copper. It is extracted from different reservoirs in the world, and most of the reservoirs are located in the Middle East. Refineries are highly complex, capital-intensive and integrated industrial plants, in which crude oil is separated and transformed into more valuable products such as LPG, gasoline, kerosene, jet fuel, diesel, asphalt, fuel oil and coke.

There is large number of refineries all over the world, and each refinery has a unique configuration and operating characteristic. However, all refineries have common chemical processes, such as desalting, distillation, hydrotreating, reforming, cracking, alkylation, isomerization and polymerization in different process units: CDU, CCR, DHP, Isomerization Unit, LPG-Amine Treating Unit. Figure 3.1, shows a schematic representation of process units in a refinery.

CDU, also called as atmospheric distillation unit, is one of the most important units in the refinery, since it affects all the downstream refining process units. In this unit, crude oil is separated into fractions with respect to their boiling points under atmospheric pressure. Products are LPG, naphta, kerosene, diesel and atmospheric residue, which are sent to the downstream units for further treatment. Atmospheric residue is sent to the Vacuum Distillation Unit (VDU), from which light vacuum gas oil (LVGO), heavy vacuum gas oil (HVGO), and asphalt is obtained. Hydrocracking Unit is operated at high pressure, and hydrogen is used for the catalytic cracking process, by which HVGO from

VDU is converted to LPG, HSRN, kerosene and diesel products. In this unit, sulfur and aromatic contents in hydrocracked streams are also eliminated. In CCR, heavy naphtha streams are processed. With catalytic reactions, inverting hydrocarbons to aromatics increases the octane number of naphtha streams. The product, called reformate, goes to the gasoline blending. In addition, reformers produce hydrogen as side product and it is sent to the refinery hydrogen network to be used in refinery processes. In Isomerization Unit, low-octane C5- C6 paraffin molecules are rearranged to form high-octane C5-C6 iso-paraffins, in order to produce high-quality Light Straight-Run Naphtha (LSRN) gasoline blend stock [40]. The product of this unit is called isomerate, which does not include any sulfur or benzene. In DHP, LD, HD from crude distillation and LVGO from vacuum distillation are used as feedstock. In this unit, components such as sulfur, and nitrogen are removed via hydrotreating, i.e. catalytic reactions under hydrogen environment, and diesel product is obtained.

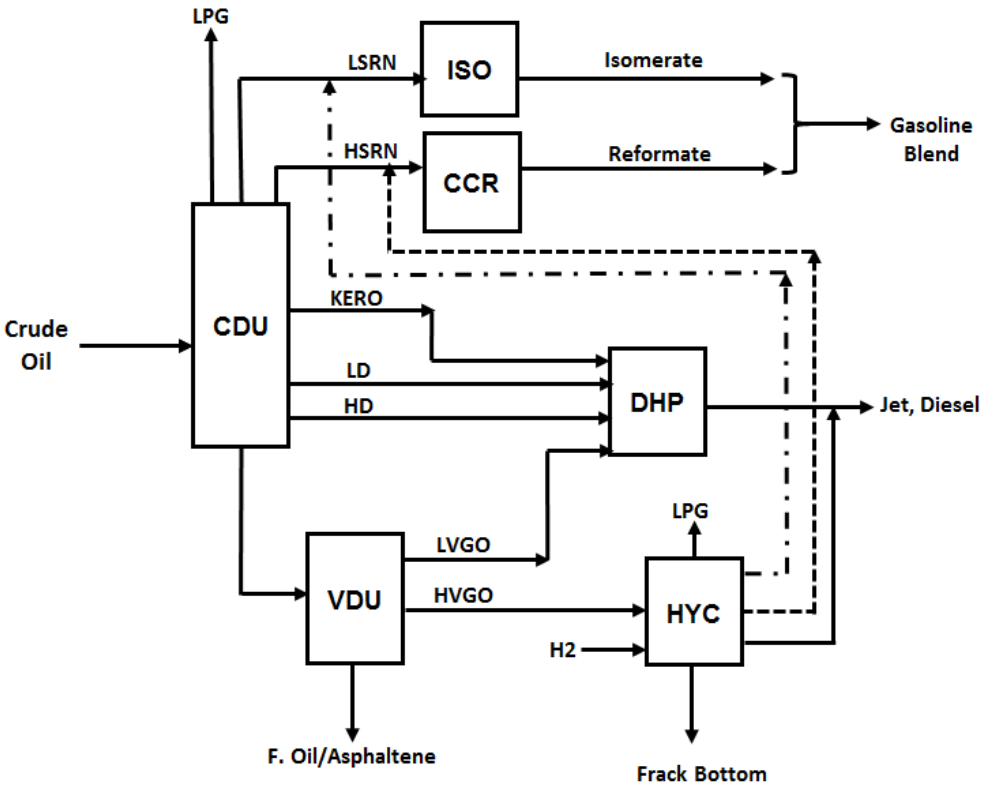

Figure 3.1. General process schema of refinery.

In refineries, beside process units, there are utilities such as hydrogen production, wastewater treatment, electricity generation, steam generation and fuel gas recovery systems [40]. By this way, refineries reduce dependency to external sources.

The primary economic objective of the refineries is to maximize the value added to the final products to make the properties of final products, such as octane level, sulfur content, and T95 value, comply with those in the sales spectrum [40]. Objective of refineries may change for different regions of the world. For example, most of the refineries aim to maximize gasoline production in North America, while refineries in the other regions of the world mostly aim to maximize diesel production, due to growing demand in the world [40].

3.2. Distillation Theory

Distillation is an essential process for refineries and almost all process units include distillation section. Distillation is a method of separation for purifying liquid mixtures by maintaining vapor and liquid phases at essentially the same temperature and pressure as coexisting zones. As the system moves toward equilibrium, each species in the mixture attains a different concentration value in each zone [42]. There are various types of distillation columns equipped with trays, or packing to provide interphase for vapor and liquid phases. Depending on the process, distillation may be batch or continuous, and refinery processes are generally continuous processes.

Distillation is usually grouped into two: simple distillation and fractional distillation. In simple distillation, two liquids having different boiling points are separated. For this purpose, mixture is heated to carry the volatile components up at the entrance of the column. Mixture is separated into two products with the help of the condenser and reboiler, Fractional distillation is used, when boiling points of the mixture components are very close to each other.

In Figure 3.2, a typical two-product distillation column is shown. The feed entering the column is separated into fractions. Due to density difference, after the feed enters the column, it is flashed and liquid phase runs down, while vapor phase flows up contacting to

each tray [42]. With respect to the feed location, the upper part of the column is called “rectifying” section, while the bottom part is named “stripping” section. In the rectifying section, vapor at the top of the column is condensed in the condenser, and some of this condensed liquid is given back to the column as reflux to provide overflow, while the rest is withdrawn as the distillate. In the stripping section, some part of the liquid running down to the bottom is sent to the reboiler and vaporized to provide the boil-up, which is sent back up the column. The remaining part is withdrawn from the bottom of the column as the bottom product [42].

3.3. Crude Distillation Process in TUPRAS

The CDU is the first process unit of the petroleum refinery, in which raw crude oil is separated into various fractions of different boiling ranges, and each of these fractions is then processed in other parts of the refinery. The CDU is composed of atmospheric distillation column, naphtha splitter column, debutanizer column and vacuum column.

In the CDU of TUPRAS İzmit Refinery, crude oil coming from the storage tanks is sent to preheat exchangers, in which crude oil is passed against hot process streams. Preheated crude is sent to the desalter, in which salt content in the crude oil is removed to prevent corrosion in piping systems and equipment (top row in Figure 3.3). Following the desalter, crude oil is passed through a second group of preheat exchangers, and then sent to furnaces, in which they are heated to the required distillation temperature. The charge leaving the furnace is fed to the atmospheric distillation column.

In the atmospheric distillation column, crude oil is separated into Naphtha, Kerosene, LD, HD and Atmospheric residue. Steam is given to the column from the bottom to enhance separation by decreasing vapor pressure in the column [44]. Top stream of the column is sent to naphta splitter column, in which Heavy Straight Run Naphta (HSRN) is drawn from the bottom of the column to be sent to the Naphta Hydrotreater and Reformer Units, and the top product is sent to the Debutanizer column to be separated as LPG and LSRN. LSRN is sent to Isomerization Unit and LPG goes to LPG treatment. Atmospheric Residue leaving from the bottom of the atmospheric distillation column is sent to the VDU, in which atmospheric residue is separated to LVGO, HVGO and Fuel Oil under vacuum.

Side cuts Kerosene, LD and HD leaving from the atmospheric distillation column are sent to the side strippers. Here, by injecting stripping steam, the flash points of the products are fixed. Kerosene, LD and HD leaving the strippers are sent to the storage, or to other process units as feedstock. LD and HD drawn from the column are sent to DHP, which has hydrotreating and hydrocracking capabilities. Also, LVGO leaving the vacuum column is sent to DHP. For this unit, planning department determined a single target value for the mixed feed T95 value, limited by catalyst operation requirements. So, independent of the individual boiling points of LD, HD and LVGO, the unit is deemed to be optimized, when T95 of their mixture is reached.

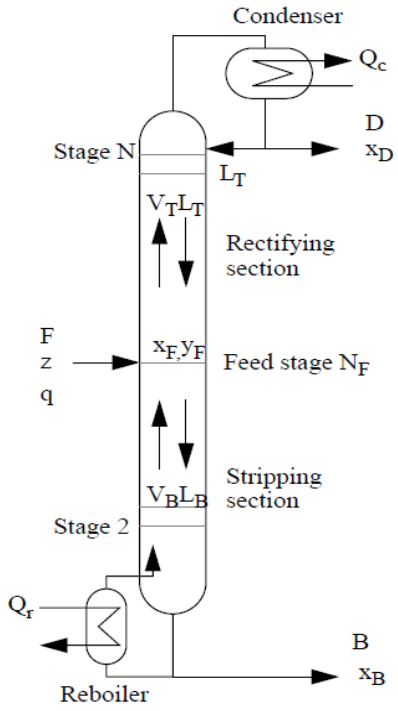

Figure 3.2. Distillation column assembly [43].

In CDU of TUPRAS İzmit Refinery, there are a number of variables, which carry high importance for the quality of the refinery products: gravity, color, distillation temperatures, boiling points, flash point, etc. Since there are no online distillation analyzers installed in CDU, these variables are usually analyzed once in a day in the laboratories of TUPRAS.

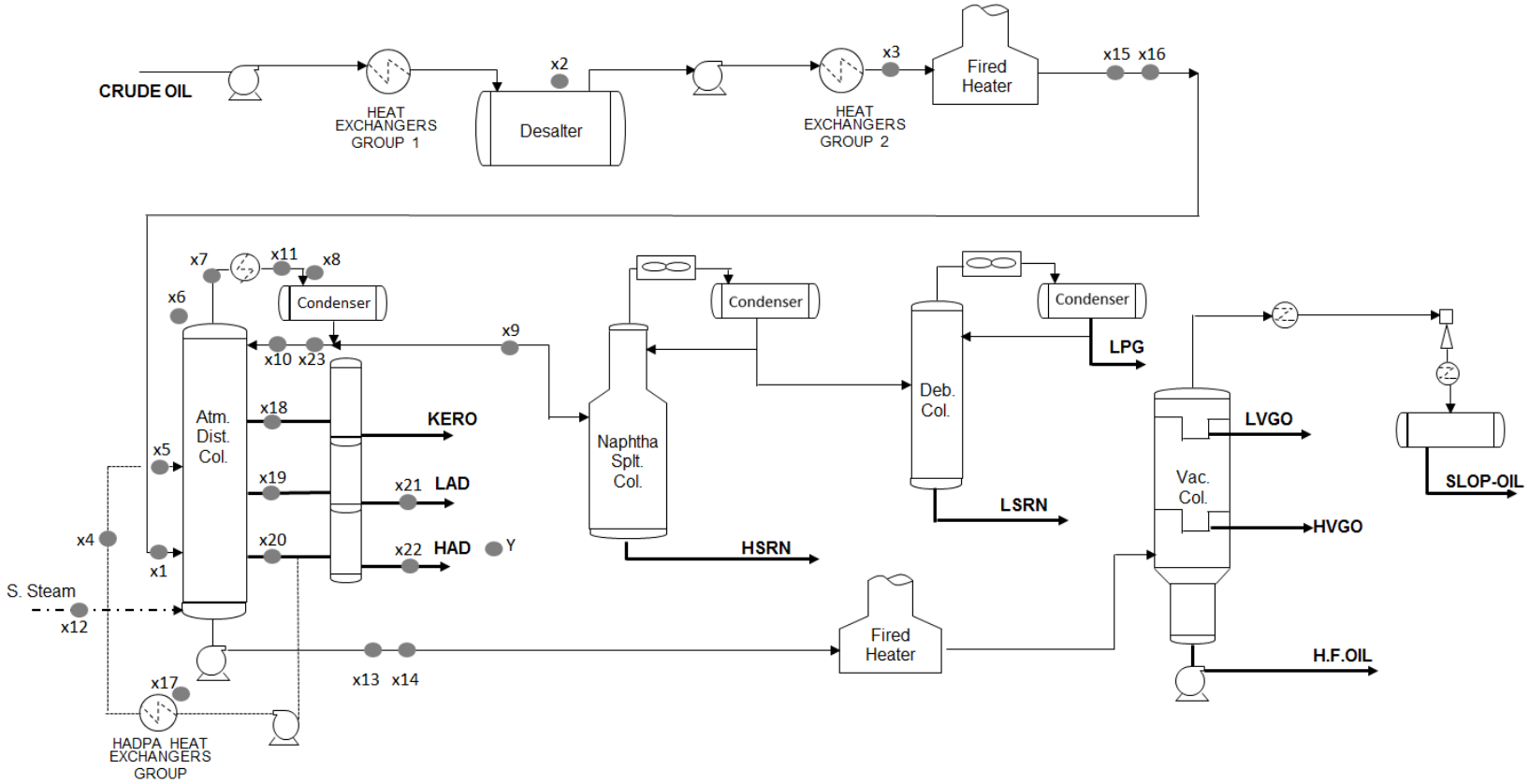

Figure 3.3. Flowchart of the CDU in TUPRAS İzmit Refinery

3.4. Laboratory Test Method for HD T95 Measurements

Boiling ranges of hydrocarbons give information about their composition and properties. Hence, distillation temperatures of refinery products are important indicators of the safety and efficiency of operation. The distillation temperature ranges of products, which are also called as distillation limits, are determined by planning department and process units with respect to the optimum, efficient operation conditions and sales demands. In order to keep required distillation temperature of product within the limits, process operation is interfered.

In the CDU of TUPRAS İzmit Refinery, two diesel cuts are fractionated as LD and HD between kerosene and residuum. For the diesel products, T95 value is one of the most important variables related with the refinery profit. T95 represents the temperature at which 95% of diesel by volume would vaporize. Decrease in HD T95 represents an undesirable contribution of LD to HD products, while increase in HD T95 represents an undesirable contribution of residual to HD products. Properties of catalysts used in DHP reactors, i.e. catalyst life, amount of coke on catalyst, play highly important roles in determining the distillation limits of HD T95. The other important variable is the sales specifications of diesel product. Target HD T95 value is maintained at a constant value, unless there is a disturbance in the type of crude oil. Hence, keeping HD T95 within an optimum range is very important for the efficiency and profitability of the operation.

HD T95 value of the HD stream (shown with Y for the HAD stream in Figure 3.3) is analyzed in the laboratory once in a day using ASTM-D86 method. This method is used to determine boiling range characteristics of products like light and middle distillates (naphtha, kerosene, and diesel) quantitatively using distillation [45]. In Figure 3.4, representative ASTM D86 distillation curves for various products are shown. In ASTM-D86 method, sample may be classified into four groups, with respect to composition, vapor pressure, initial and final boiling points. The group characteristics are shown in Table 3.1. According to Table 3.1, HD is included in Group 4.

In ASTM-D86 method, sample taken from process units is distilled in a laboratory batch distillation unit under desired conditions to reflect reality. During the test,

temperature and volume of the distilled sample are recorded, measured temperature values are corrected via barometric pressure. If any of the desired conditions is not met, the test is repeated. Finally, test results are reported as percent evaporated, or percent recovered versus corresponding temperature [46].

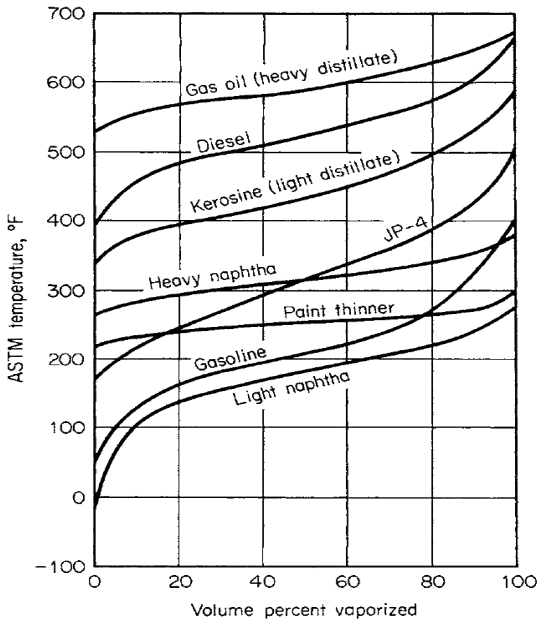

Figure 3.4. Representative ASTM D 86 distillation curves [42].

Table 3.1. Groups according to sample characteristics [46].

	Group 1	Group 2	Group 3	Group 4
Sample characteristics Distillate type Vapor pressure at 37.8 ^{0}C, kPa	≥ 65.5	< 65.5	< 65.5	< 65.5
100 ^{0}F, psi	≥ 9.5	< 9.5	< 9.5	< 9.5
(Test Methods D323, D4953, D5190, D5191, D5842, IP 60 or IP 394) Distillation, IBP ^{0}C			≤ 100	> 100
^{0}F			≤ 212	> 212
EP ^{0}C	≤ 250	≤ 250	> 250	> 250
^{0}F	≤ 482	≤ 482	> 482	> 482

Repeatability and reproducibility are two important parameters in laboratory measurements. Repeatability is the difference between successive test results, obtained by the same operator using the same apparatus under constant operating conditions on

identical test material. Reproducibility is the difference between two independent test results, obtained by different operators working in different laboratories on identical test material [46]. Table 3.2 shows repeatability and reproducibility relations for Group 4 samples. In the given relations, *T* is the temperature at which the reported volume percent of the sample is distilled. For HD produced in TUPRAS İzmit refinery CDU, according to Table 3.2, repeatability of the HD T95 value changes between 3.15 and 3.9 ^{0}C, while reproducibility changes between 8.48 and 10.9 ^{0}C.

Table 3.2. Repeatability and Reproducibility for Group 4 [46].

Percent Recovered	Repeatability ^{0}C	Reproducibility ^{0}C	Valid Range ^{0}C
IBP	0.02T	0.055T	145 to 220
10 %	0.009T	0.022T	160 to 265
50 %	1.0T	3.0	170 to 295
90 %	0.004T	0.015T	180 to 340
95 %	0.015(T-140)	0.042(T-140)	260 to 340
FBP	2.2	7.1	195 to 365

4. HISTORICAL DATA COLLECTION

For the CDU in TUPRAS İzmit Refinery, 23 process variables, which are deemed to affect HD T95 value, are selected. Locations of these measured variables are shown in Figure 3.3, and the designations of the variables are shown in Table 4.1. TUPRAS historical database was used to obtain the trajectories of the process variables for a one-year period. The laboratory measures HD T95 value once in a day; hence, process variable measurements were averaged over 4 hours about the laboratory sampling times to filter noise and make the process variable measurements more faithfully represent the process conditions, under which laboratory quality measurements are performed. The total number of collected observations is 323, and these observations were divided into two sets: the first 200 observations were used for constructing the models (training), while the remaining 123 data points were used for testing the model. Figure 4.1 shows the individual time trajectories of the process variables used for training, with the 99% confidence limits (CLs) shown with dashed lines, obtained via normal distribution assumption of the data.

Table 4.1. CDU process variables used for modeling.

Process Variable Designation	**Explanation**
x_1	Crude charge flow
x_2	Desalter pressure
x_3	2nd group heat exchangers exit temperature
x_4	HD reflux flow
x_5	Column exit temperature of HD reflux
x_6	Column pressure
x_7	Column top temperature
x_8	Condenser drum temperature
x_9	Flow to Naphtha Splitter column
x_{10}	Top reflux temperature
x_{11}	Air cooler exit temperature
x_{12}	Stripping Steam temperature

Process Variable Designation	Explanation
x_{13}	Column bottom flow 1
x_{14}	Column bottom flow 2
x_{15}	Fired heater transfer temperature 1
x_{16}	Fired heater transfer temperature 2
x_{17}	HD reflux reboiler exit temperature
x_{18}	Kerosene temperature
x_{19}	LD temperature
x_{20}	HD temperature
x_{21}	LD flow
x_{22}	HD flow
x_{23}	Top reflux flow
Y	HD T95

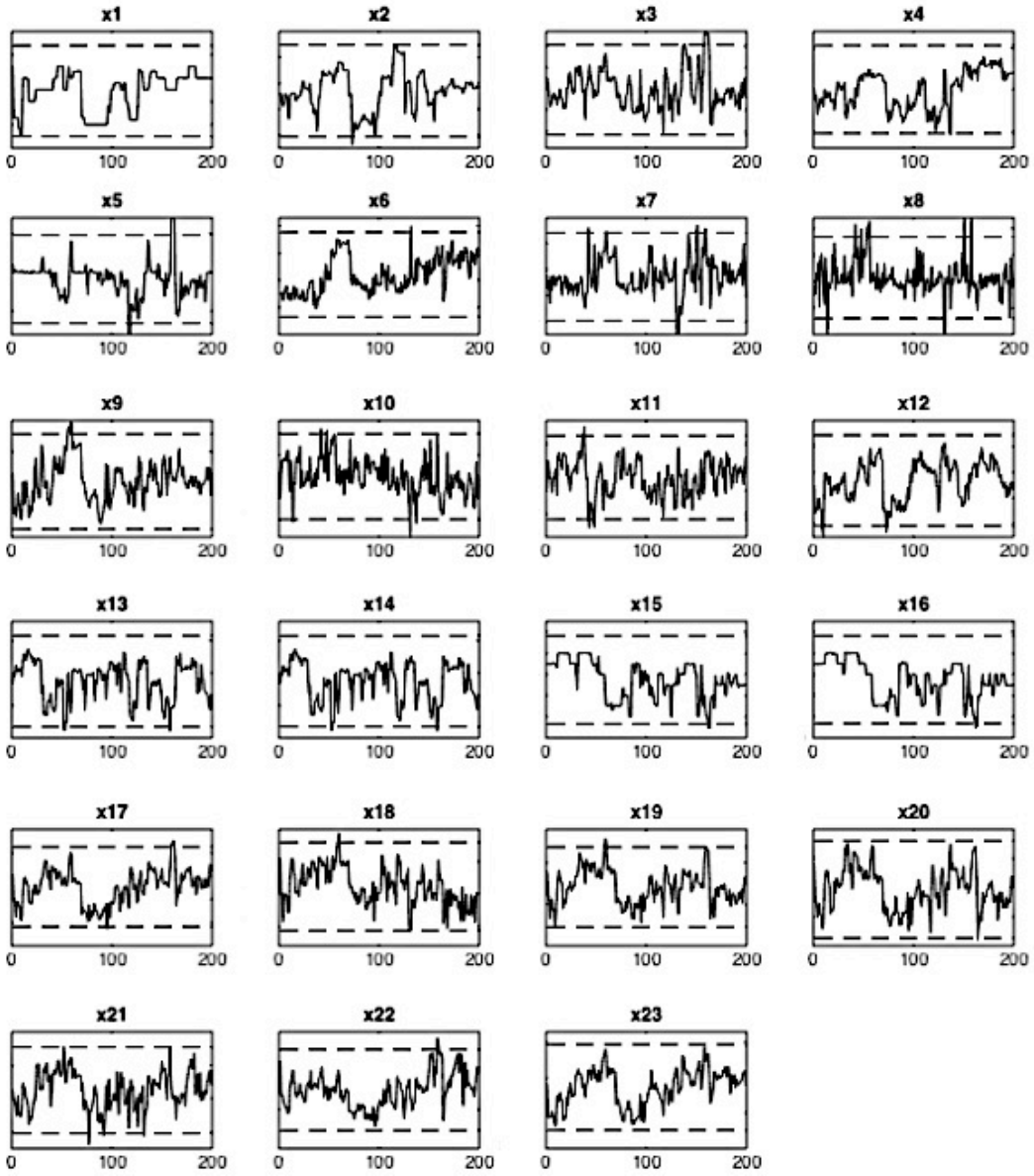

Figure 4.1. Time trajectories of CDU process variables in the training set.

5. RESULTS AND DISCUSSION

Results section is divided into three parts: Monitoring historical operation (Section 5.1), MC simulations (Section 5.2) and model prediction (Section 5.3). In Section 5.1, application of PCA and MCD methods on CDU process to identify the relationship between process variables are described in detail. In Section 5.2, evaluation of predictive performances of LS and various robust regression methods using MC simulations are explained. In Section 5.3, application of LS and robust regression methods to predict HD T95 in detail.

5.1. Monitoring Historical Operation

In this section, it is aimed to determine relations between the 23 processes variables used in modeling the CDU. It is also aimed to determine a convenient and efficient method for monitoring CDU. For this purpose, PCA (see Section 2.2) was applied on the process dataset to extract the essential operating conditions and reduce the dimensions of the process, while robust MCD (see Section 2.3.5) method was employed to prevent the detrimental effect of outliers on the identification of the covariance structure between the variables.

In monitoring CDU, two different methods were used. First, PCA was applied to the process variables, and observations exceeding the 99% CLs were removed from the dataset and PCA was employed on the reduced dataset. The resulting model was called as “skipped-PCA” in the current study. Second, MCD with 25% BDP was applied to the process variables. Outliers identified by 99% CLs in MCD were removed from the dataset and PCA was employed on the remaining data. Like that in the first method, observations exceeding 99% CL were removed, and PCA model was reconstructed for the remaining data. This model will be called as “MCD+PCA” in the current study. PCA functions and MCD functions (FSDA-Toolbox) in MATLAB were used for calculations [47].

5.1.1. Application of the Skipped-PCA Method on Historical Data

After auto scaling, i.e. subtracting each variable from its mean and dividing to its standard deviation, PCA was employed on the 200 observations from 23 process variables. In PCA, it is important to know how many principal components (PCs) should be used in order to account for most of the data variability [48]. This quantity was found by applying cross-validation (CV) method, which yields prediction residual sum of squares (PRESS) for different number of PCs. PRESS residuals from a model represents the model's ability of prediction, and smaller PRESS value indicates a better prediction performance. The optimum number of PCs is usually chosen to be the first minimum point in the PRESS residuals profile. It is seen in Figure 5.1 that six PCs has the smallest PRESS residuals from CV analysis.

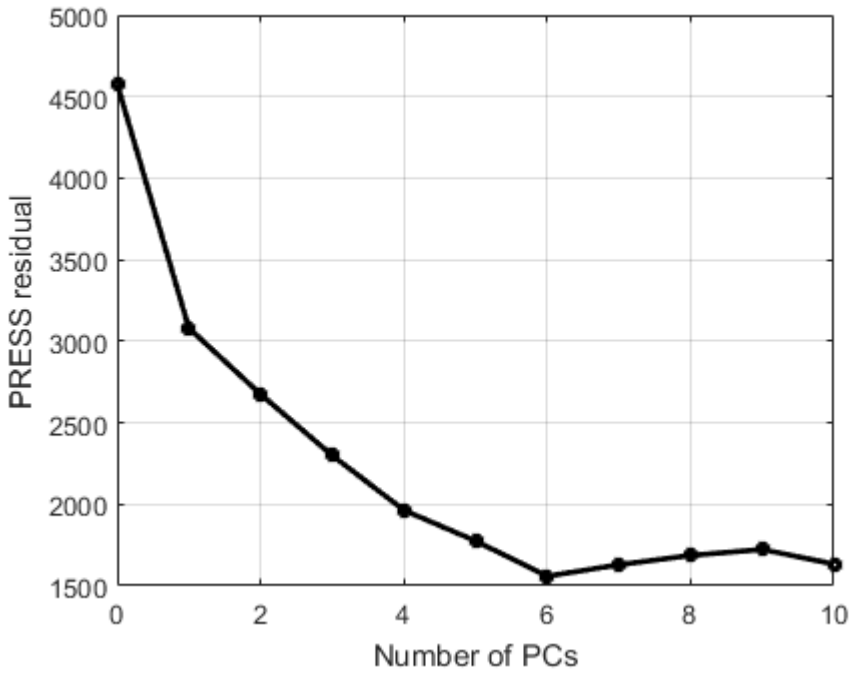

Figure 5.1. PRESS residuals vs. number of PCs.

Percentage and the cumulative percentage explanations of variables obtained via PCA are shown in Figure 5.2. In Figure 5.2-a, dashed lines represent the explanation percentage when process variables are randomly and independently distributed, and intersection of both lines is another indicator for the optimum number of PCs. Here,

similar to that found by CV, optimum PC number is found to be six. Cumulative percentage explanation plot shows that the six-PC model explains ~83% of variation in the process data.

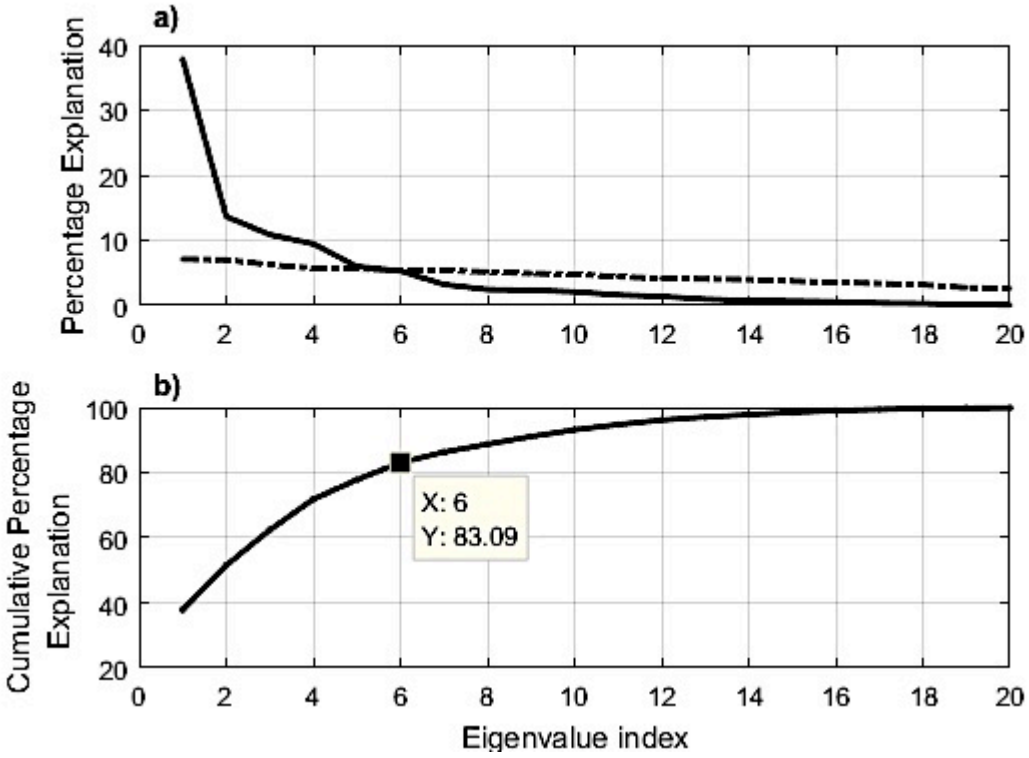

Figure 5.2. (a) Percentage explanation of PCs and (b) Cumulative percentage explanation of PCs.

Next a PCA model with six PCs was constructed. In Figure 5.3, contribution of process variables to each PC is showed. Colors from darkest to lightest match with the indices of PCs, i.e. the darkest color correspond to first PC, while the lightest color corresponds to the sixth PC. It is seen that variables x_3, x_9, x_{17}, x_{19}, x_{20} and x_{23} are highly correlated with PC 1, while variables x_{12}, x_{13} and x_{14} dominate PC 2.

Figure 5.4 shows the two-dimensional distribution of t-scores, which are the projections of the data in the reduced PC spaces. It is seen that the bivariate distributions are somewhat like multivariate normal distribution. There are not any samples out of CLs, representing different operation condition. On the other hand, it is also possible that

outliers may be masked by the presence of a group of samples, which come from a different distribution than the one representing normal operation data.

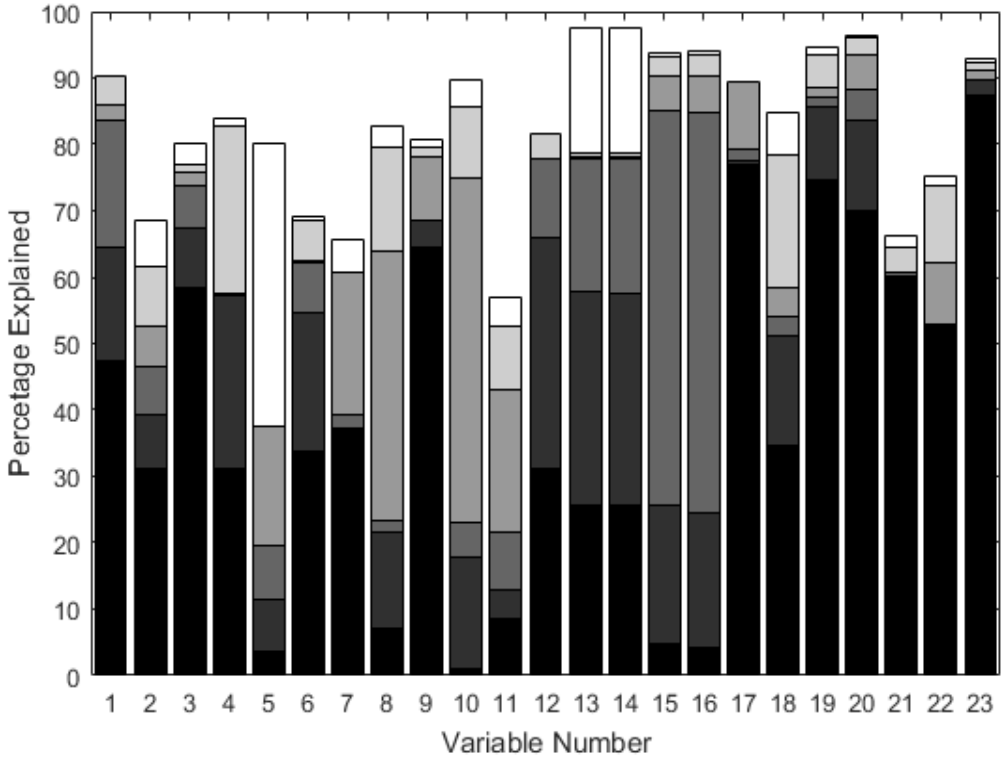

Figure 5.3. Percentage explanation of variables in each PC.

The T^2 and Q-residual measures are commonly used statistics for PCA diagnostics. As mentioned in Section 2.2, the Hotteling's T^2 measures the variation within the PCA model, e.g. a high T^2 value shows that process is currently in an excessive operation point. On the other hand, the Q-residual measures the lack of model fit for each sample, e.g. a high Q-residual statistic shows that relation (covariance structure) between the variables is perturbed. The T^2 and Q-residual statistics are shown in Figure 5.5, in which the dashed lines represent 99% CLs, used to identify possible outliers. Between the 115^{th} and 140^{th} observations, there are a number of observations exceeding the 99% limit of Q-residual plot. By this method, five observations, which make up of 2.5% of the reference dataset, are deemed to be outliers (Table 5.1). Removing these outliers yielded a new dataset (dataset-1) having 195 observations. PCA was again employed on the new dataset, and the smallest PRESS residuals were obtained for six PCs (Figure 5.6), which explain %84.04 of

the dataset-1 (Figure 5.7). Percentage explanations of the variables (Figure 5.8) and distribution of the scores (Figure 5.9) have not substantially changed compared to those from the previous model (Figure 5.3 and Figure 5.4).

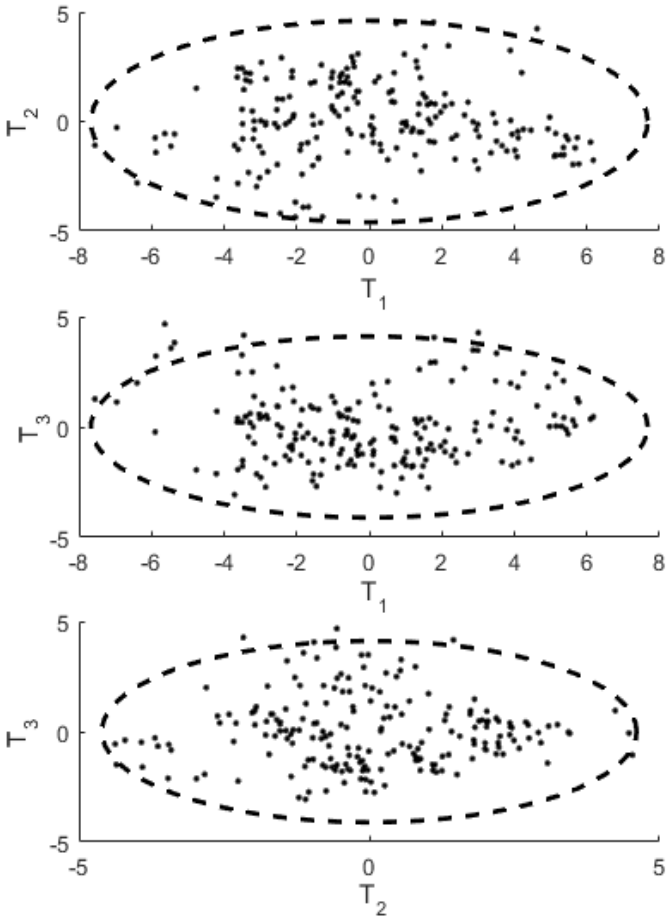

Figure 5.4. Distribution of t-scores in two-dimensional spaces.

Table 5.1. Outliers detected by PCA.

Outliers found by PCA application	118, 130, 132, 136, 137

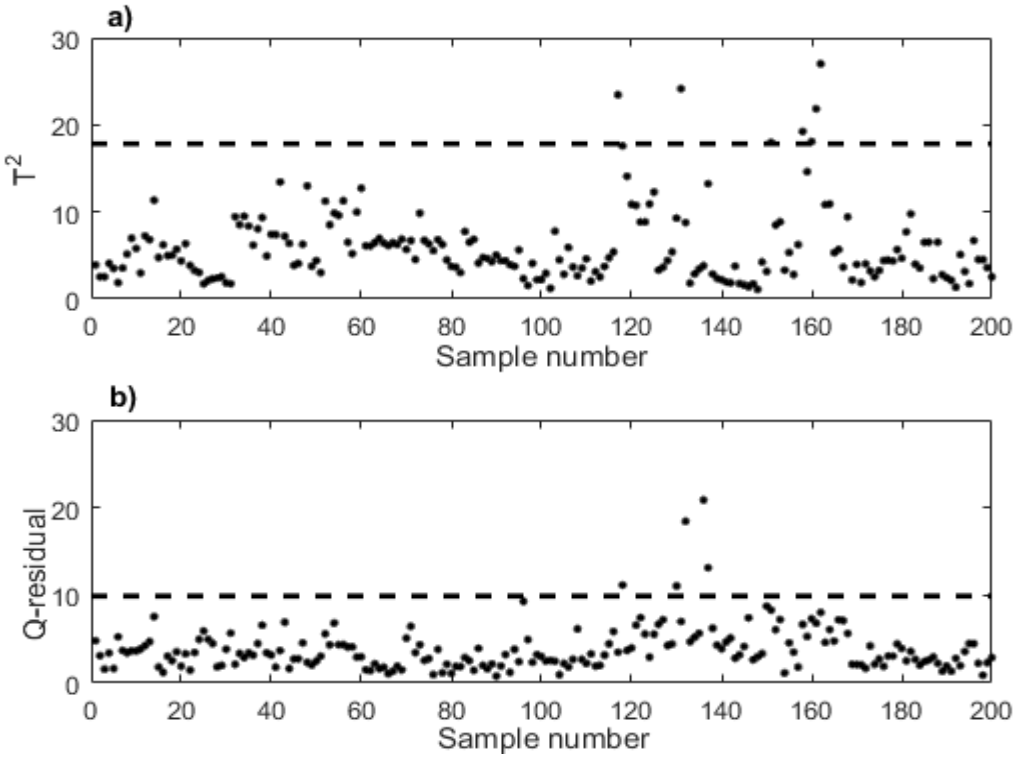

Figure 5.5. (a) T^2 statistics of the PCA model with six PCs, (b) Q statistics of the PCA model with six PCs.

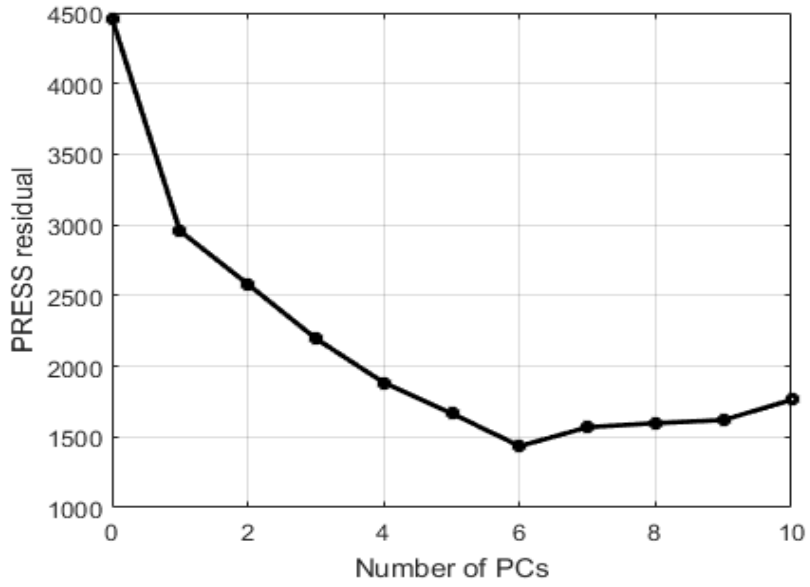

Figure 5.6. PRESS residuals vs. number of PCs in the skipped-PCA model.

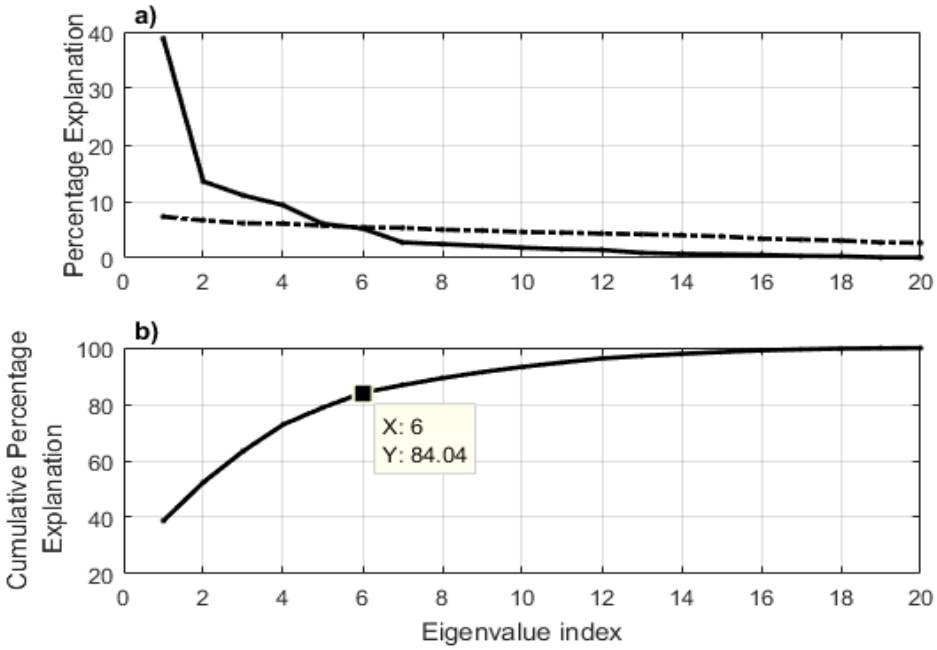

Figure 5.7. (a) Percentage explanation of new PCs and (b) Cumulative percentage explanation of PCs in the skipped-PCA model.

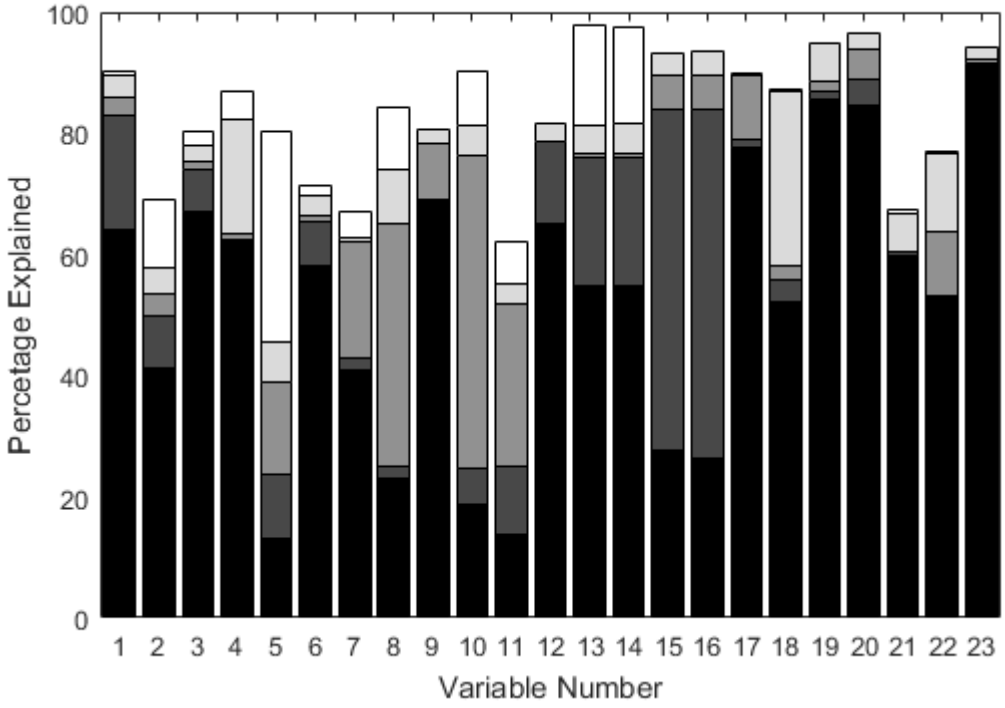

Figure 5.8. Percentage explanation of variables in PCs in the skipped-PCA model.

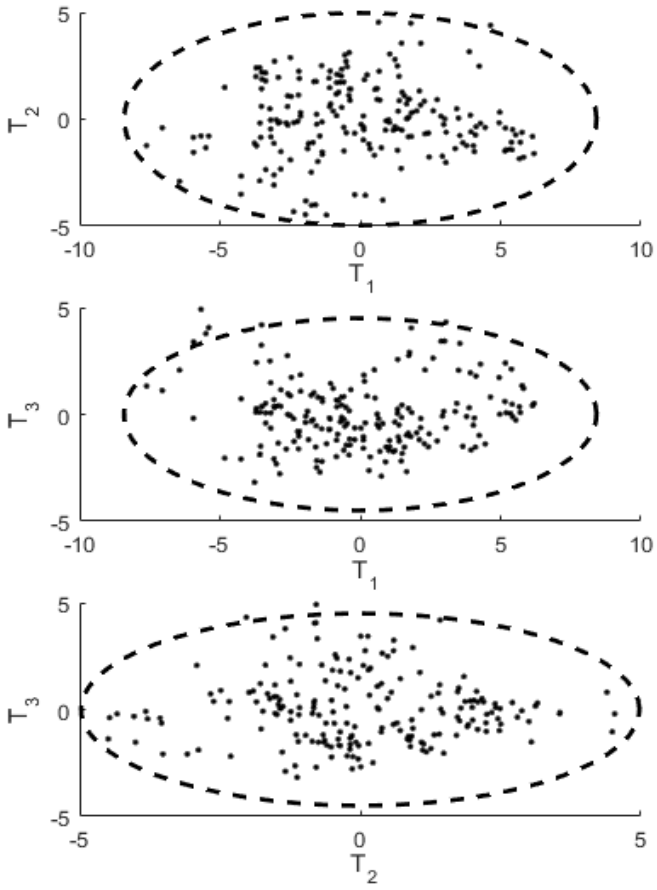

Figure 5.9. Distribution of t-scores in two-dimensional space of the skipped-PCA model.

In Figure 5.10, T^2 and Q-residual statistics of the new constructed model with the dataset-1 is shown. Q-residuals of all observations are below the 99% CL. Only, observation number 96 and 145 are approximately on the limit and observation number 129 is very close to the CL.

To validate the performance of the skipped-PCA model, 123 data points taken from the TUPRAS historical database was used as test dataset. Figure 5.11 shows the T^2 and Q-residuals statistics for the test dataset. It is seen that the last ~10 observations are significantly out of the 99% CL, showing that the process had operated out of its normal operational conditions during these time intervals. Furthermore, there are other temporary time intervals, during which relations between process variables were violated, such as observation 54 to 58, 106 to 111, and 115 to 123.

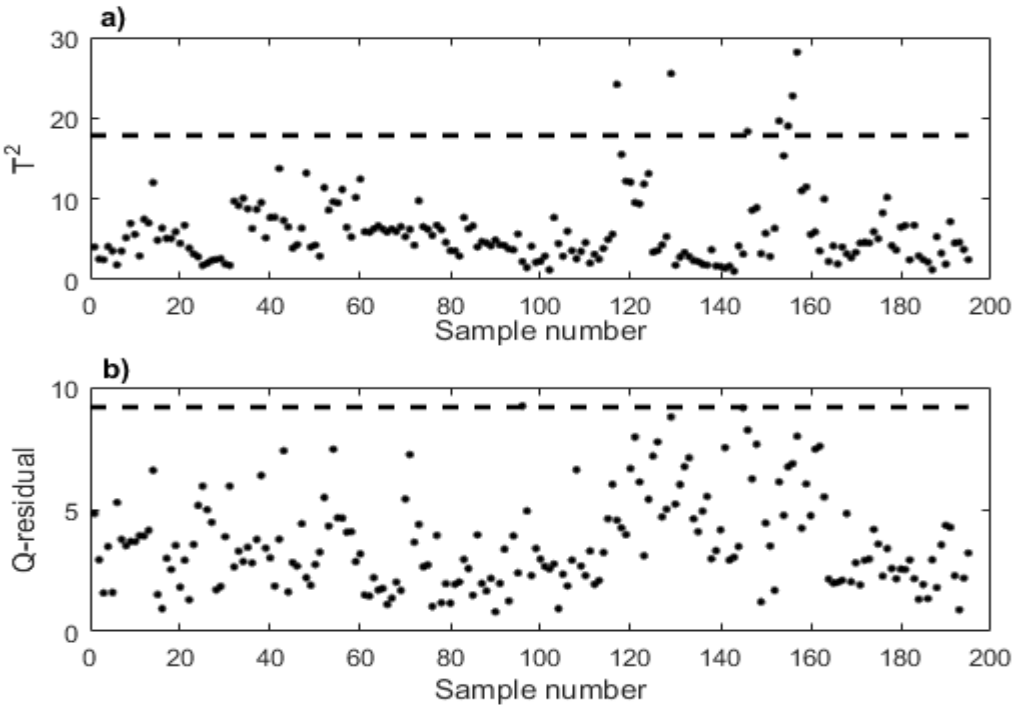

Figure 5.10 (a) T^2 statistics on the new model with six PCs, (b) Q statistics with six PCs in the skipped-PCA model.

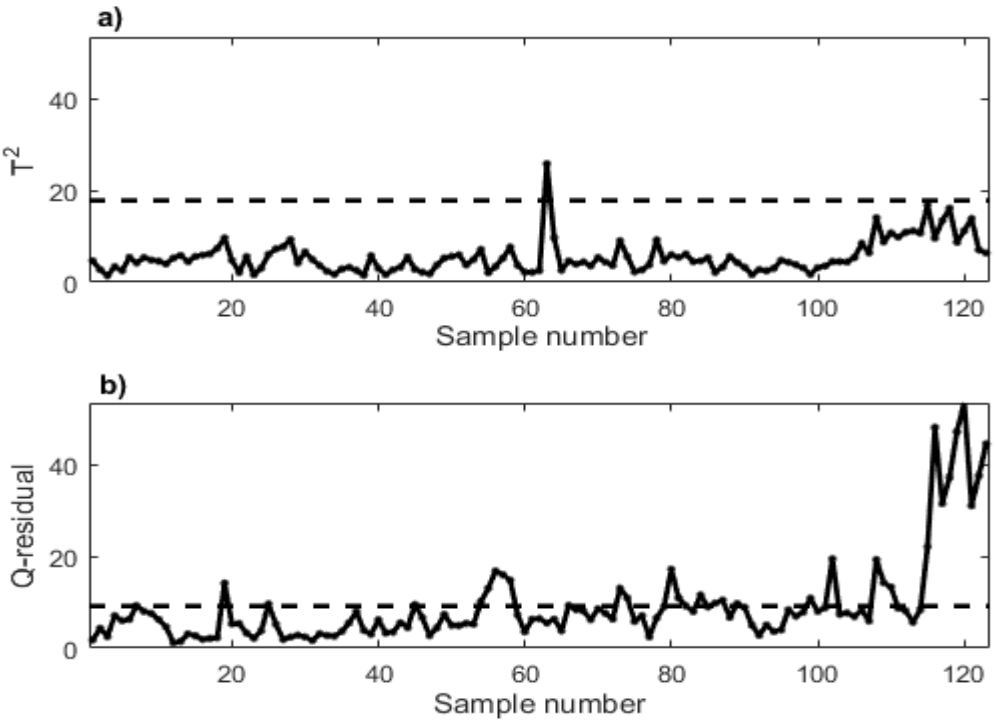

Figure 5.11. (a) T^2 statistics and (b) Q statistics on test data.

5.1.2. Application of MCD+PCA on Historical Data

The second method used in monitoring historical data involved MCD method, which is a well-known PCA robustification method. In application of this method, highly correlated variables out of 23 process variables were eliminated from the dataset, reducing the dataset to 20 variables. Eliminated variables are x_{14} (column bottom flow2), x_{16} (fired heater temperature2) and x_{20} (HD temperature). MCD was employed on 200 observations and 19 process variables. BDP and CL were selected as 25% and 99%, respectively, and MCD model was iterated for 2×10^5 times. MCD model yielded 43 outliers, comprising 20% of the reference dataset (Table 5.2). After applying MCD estimator, skipped-PCA method was followed. By eliminating identified outliers from the dataset, a new dataset (dataset-2) was constructed. Then, PCA method was applied for the dataset-2. As seen in Figure 5.12, there are possibly two minimum points indicating five PCs and seven PCs, however Krzanowski recommends five PCs to select. At this step, five PCs were selected and the model was constructed over five PCs. Figure 5.13 shows that five PCs can explain ~82% of dataset. It should be noted that MCD+PCA model yields a smaller subspace compared to skipped-PCA method.

Table 5.2. Outliers detected by MCD application.

Outliers found by MCD application	1, 14, 52, 53, 54, 55, 56, 59, 60, 96, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 130, 131, 132, 136, 137, 151, 152, 153, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168

By using the dataset-2 and 5 PCs, PCA model was constructed. Figure 5.14 shows how effectively the PC's can explain the variables. x_1, x_2, x_9, x_{17}, x_{19}, x_{20} and x_{23} have very high contribution to PC 1. It should be recalled that skipped-PCA method yielded variables x_3, x_9, x_{17}, x_{19}, x_{20} and x_{23} with high contribution to PC 1 (Figure 5.3), showing that results obtained by skipped-PCA and MCD+PCA methods are slightly but not negligibly different.

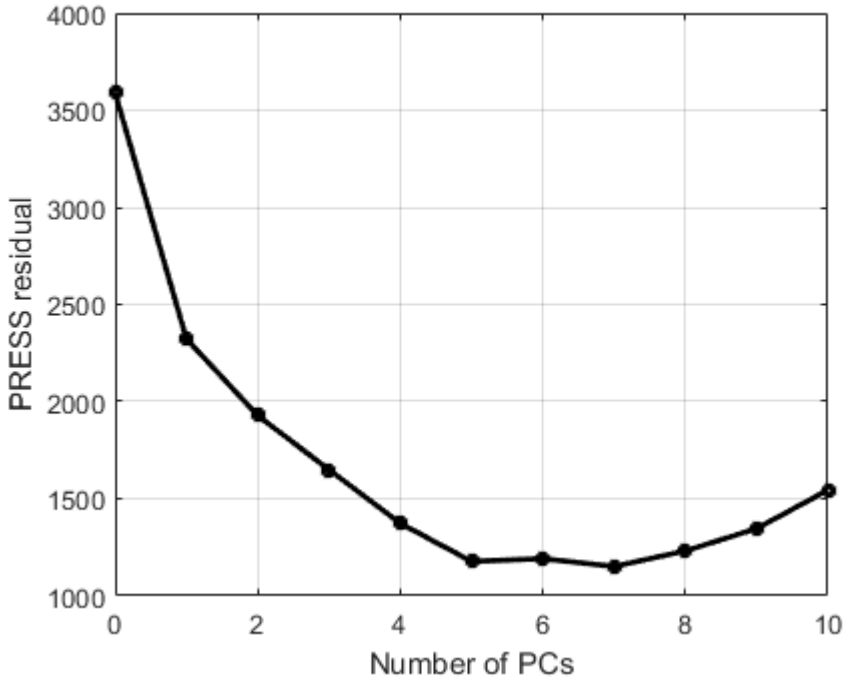

Figure 5.12. PRESS residuals vs. number of PCs of MCD based PCA method.

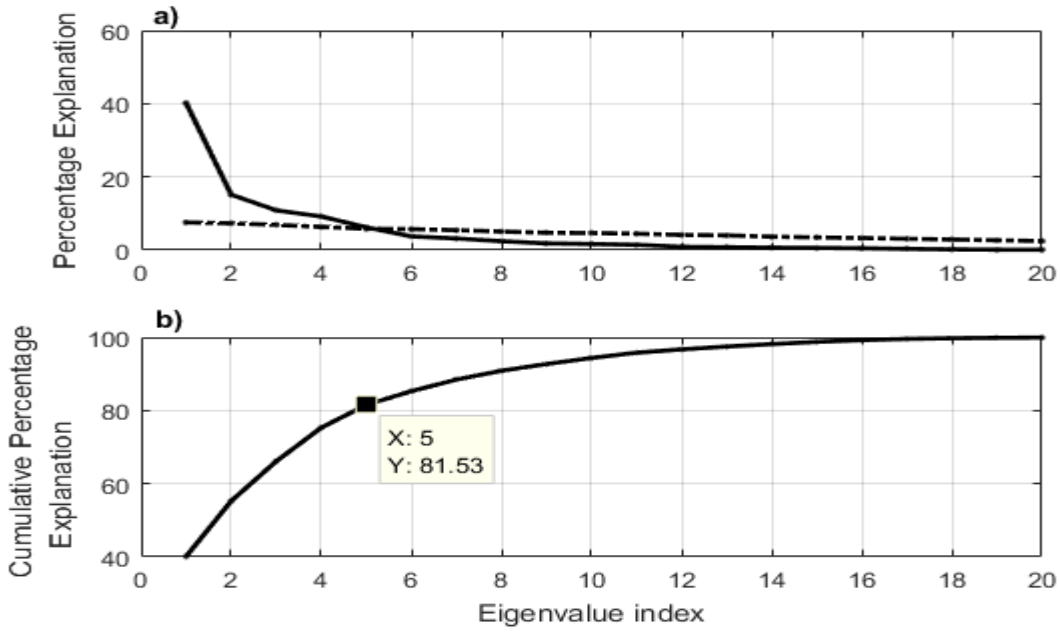

Figure 5.13. (a) Percentage explanation of PCs and (b) Cumulative percentage explanation of PCs of MCD based PCA method.

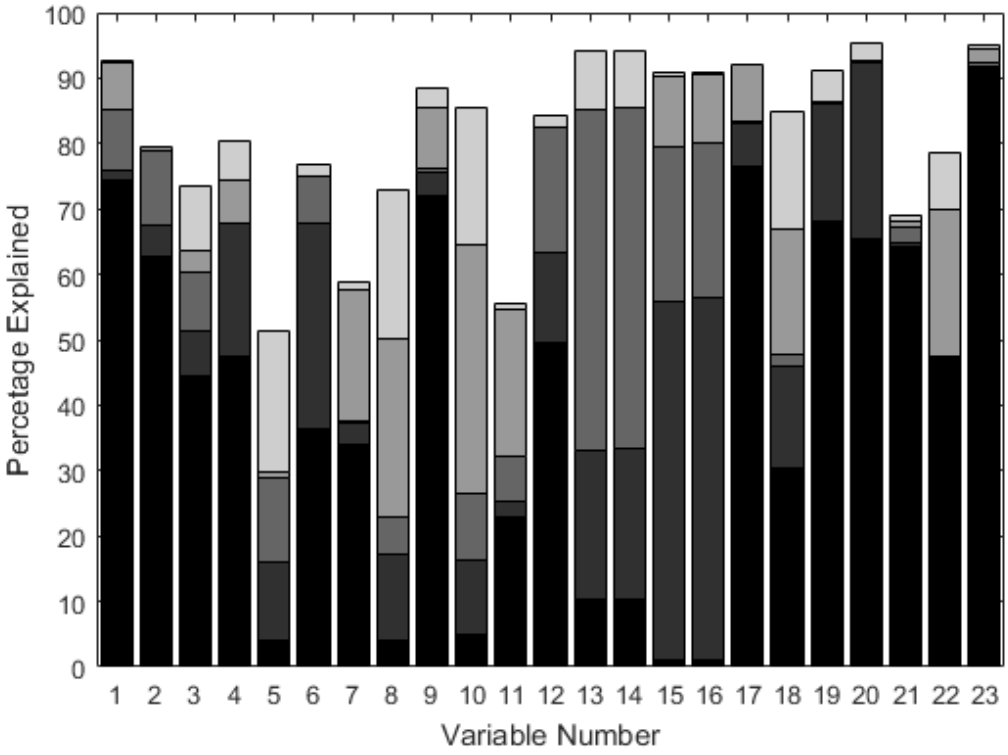

Figure 5.14. Percentage explanation of variables in PCs of MCD based PCA method.

Figure 5.15 shows the distribution of t-scores in two-dimensional PC spaces. It should be recall that there was not any evidence of perturbed operation conditions by the PCA method (Figure 5.4). However, in Figure 5.15, there is a region where data points are out of the CL in t_2-t_1 plot. It shows that MCD+PCA method can identify the perturbed process conditions more efficiently.

Figure 5.16 shows the T^2 and Q statistics of the MCD based PCA model with the dataset-2. According to Q-statistics, there are only 2 observations on the %99 CL.

The resulting MCD+PCA model was validated using the test dataset. Figure 5.17 shows the T^2 and Q-residual statistic of the test data, projected on the MCD+PCA model. Similar to that seen in Figure 5.11, there are four main time periods, during which relation between the processes variables had been changed with respect to the historical process operation: the last ~10 observations, and observations 54 to 58, 106 to 111, and 115 to123.

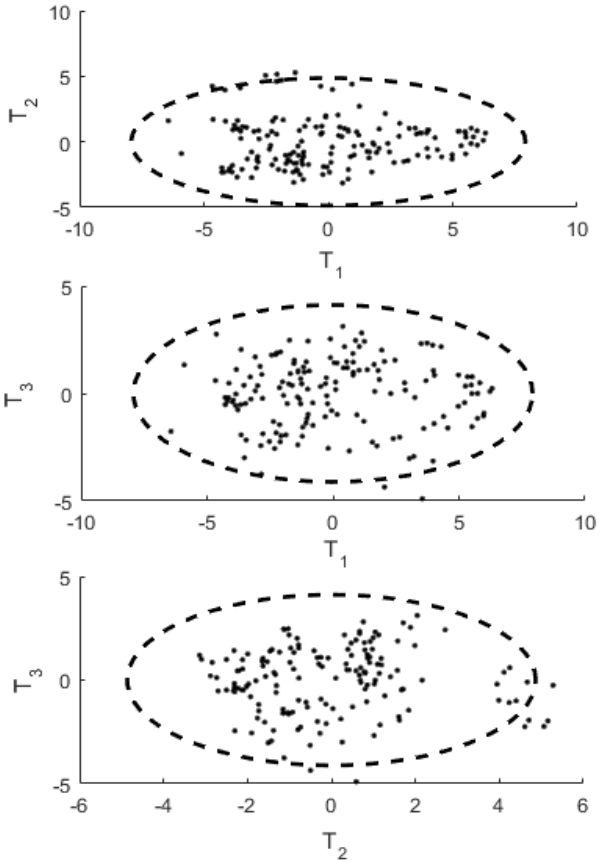

Figure 5.15. Distribution of t-scores in two-dimensional space of MCD based PCA method.

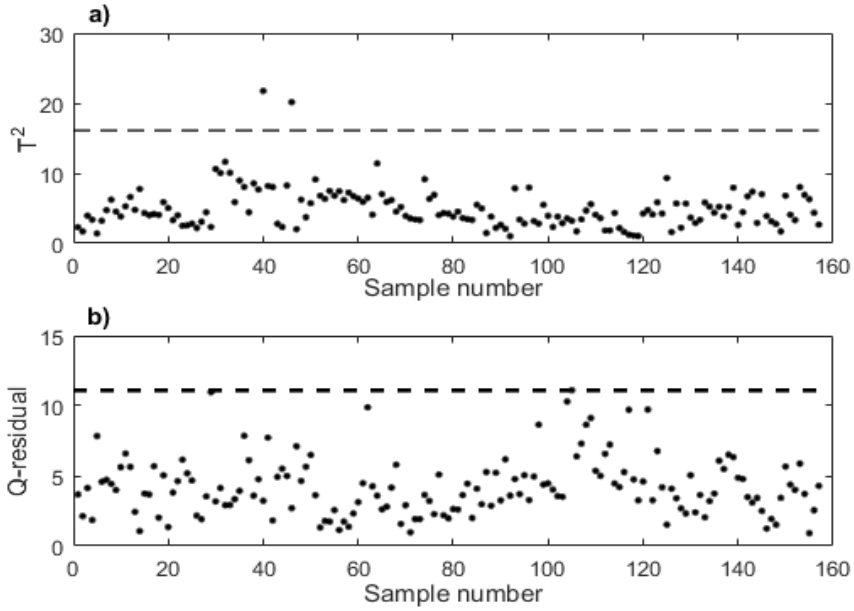

Figure 5.16. T^2 statistics and Q statistics results of MCD based PCA method.

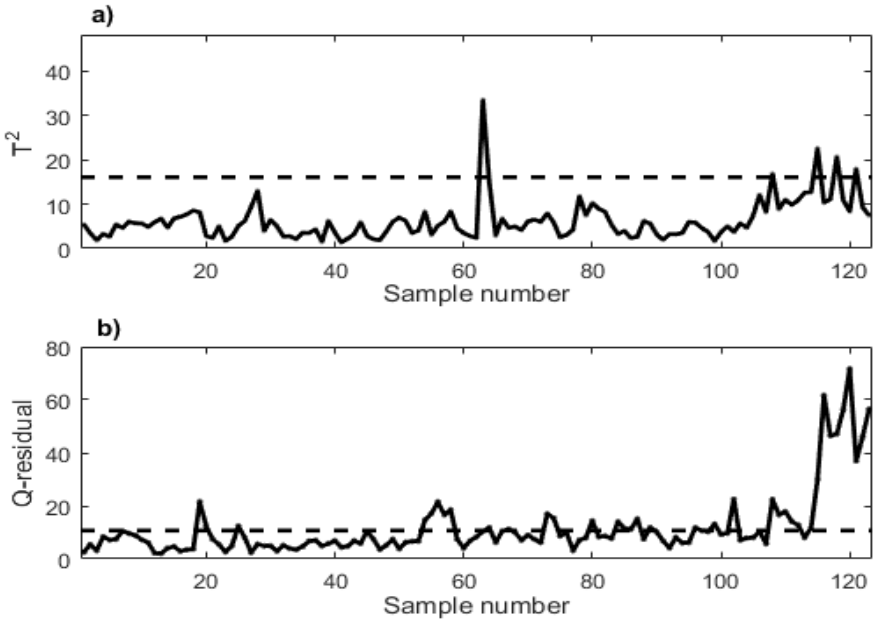

Figure 5.17. T^2 statistics and Q statistics on test data based on MCD+PCA model.

Figure 5.18 shows the Q-residual values of skipped-PCA and MCD+PCA methods. According to Figure 5.18, observations 118, 130, 132, 136, 137 are common outliers for both methods. Moreover, there are many outliers, shown at the right bottom region, detected only by the MCD+PCA model.

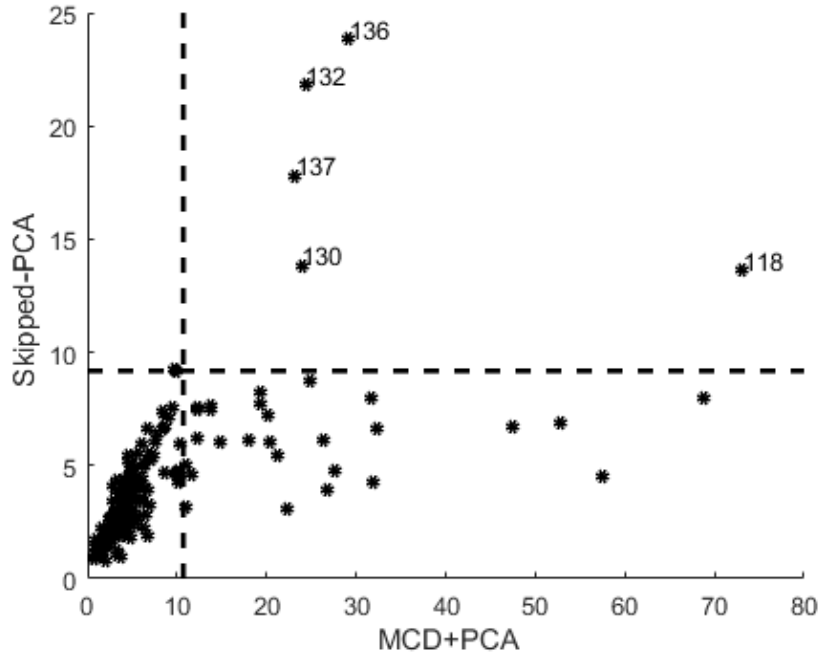

Figure 5.18. Q-residuals of skipped-PCA and MCD+PCA models.

Figure 5.19 and Figure 5.20 show the Q-residuals of skipped-PCA method and MCD+PCA models, respectively, with respect to time. In these figures, filled circles denote the outliers, which can only be detected using MCD+PCA model, i.e. these data points are outside of the 99% CL of MCD+PCA model, but inside the 99% CL determined by the skipped-PCA method. It is seen that MCD+PCA method can detect many samples between observations 120 and 170, which seem to represent process operation out of normal conditions, while only a number of incidental outliers are detected by the skipped-PCA method for the same region.

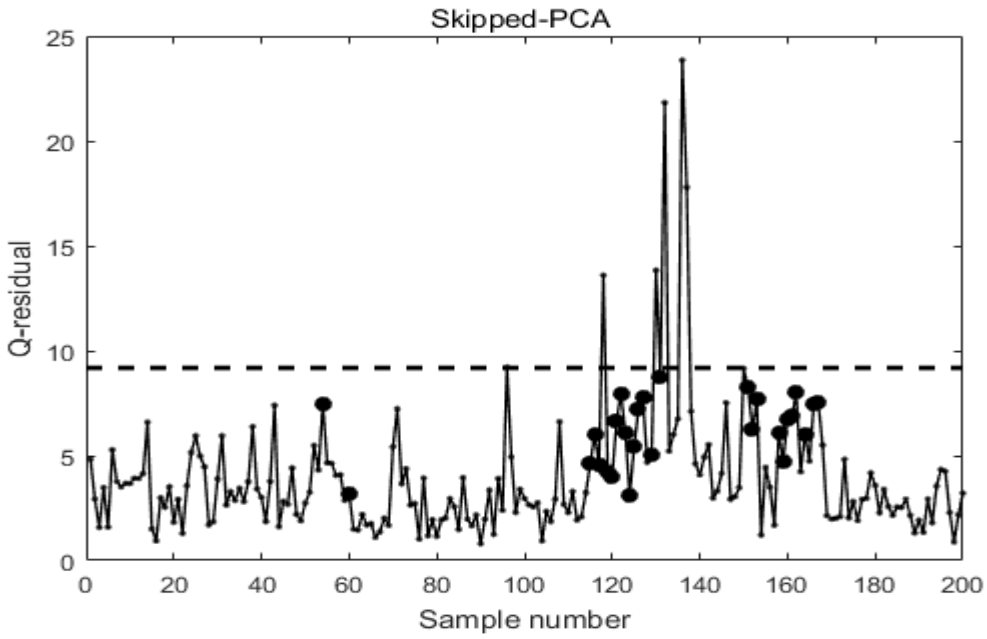

Figure 5.19. Q-residuals of the skipped-PCA method.

By using PCA for process diagnostics, process variables with the deemed perturbation were identified. The most highly disturbed process variable was found to be x_5, which is the column exit temperature of HD reflux, i.e. x_5 was found to be the single most important process variable in determining whether the process is in the normal operating range, or not. Figure 5.21 shows the relationship between x_5 and x_6. Filled data points indicate observations detected only by MCD+PCA method. As seen, x_5 tends to take values out of the boundary of the bulk of the data, and the correlation between x_5 and other process variables, such as x_6, is deteriorated. Time trajectory of x_5 is given in Figure 5.22,

where wide fluctuations in x_5 are only seen for the samples between 120 and 170, the very same time interval deemed to be problematic by MCD+PCA method.

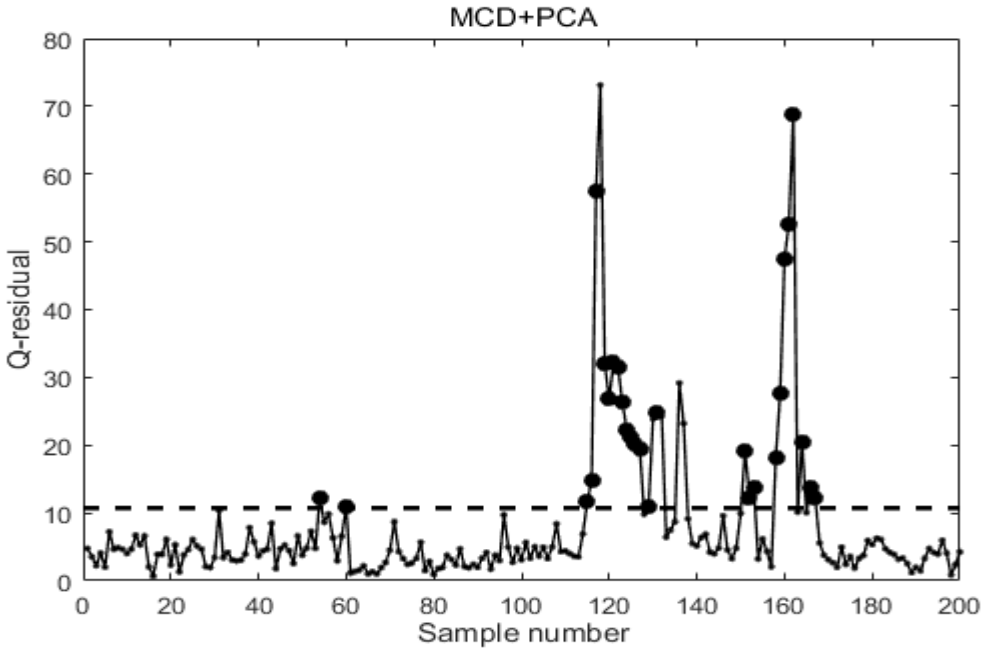

Figure 5.20. Q-residuals of the MCD+PCA method.

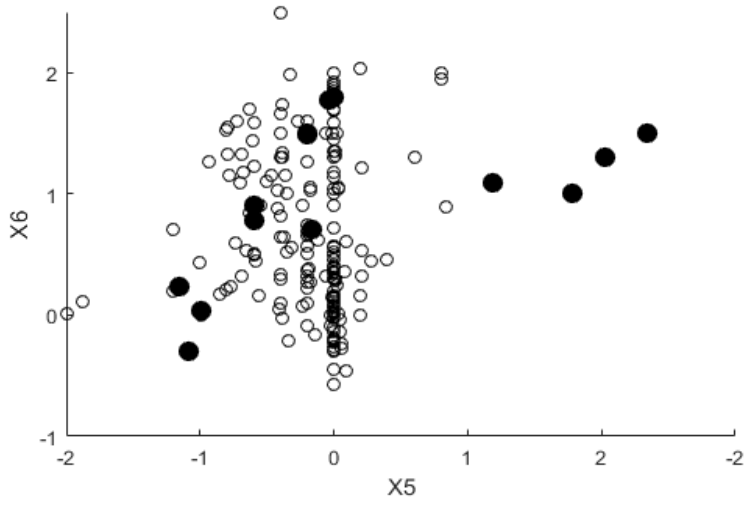

Figure 5.21. x_6 vs. x_5 for the training data.

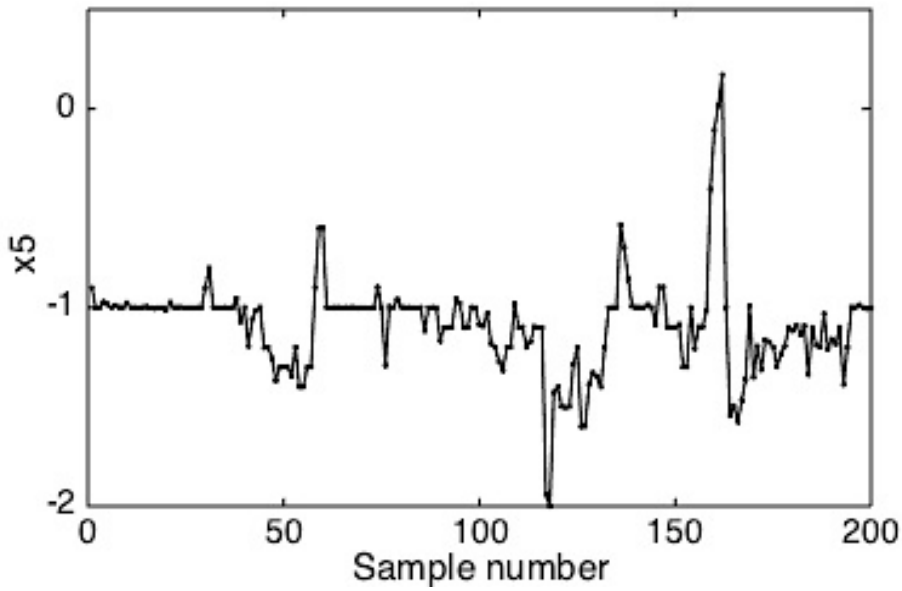

Figure 5.22. Time trajectory of x_5.

After constructing skipped-PCA and MCD+PCA methods using the reference data, test data was monitored using these two models to see how effectively abnormality in operation could be detected. Figure 5.23 shows the Q-residuals of skipped-PCA method vs. MCD+PCA method for the test data. Furthermore, Figure 5.24 and Figure 5.25 show the Q-residuals trajectories of the test data obtained by skipped-PCA and MCD+PCA models. Like that in Figure 5.19 and Figure 5.20, the filled circles denote the observations deemed to be outliers solely by the MCD+PCA method. It is seen that samples shown with filled circles are also very close to the 99% CL of the skipped-PCA model (Figure 5.24), showing that performances of MCD+PCA and skipped-PCA models in identifying out-of-control samples in test data are similar. Both in Figure 5.24 and Figure 5.25, the last 10 observations seem to be operating out of normal operational conditions. In Figure 5.26, the filled data circles represent the last 10 observations in the test data. As seen, x2 values take significantly differently values, compared with the historical data.

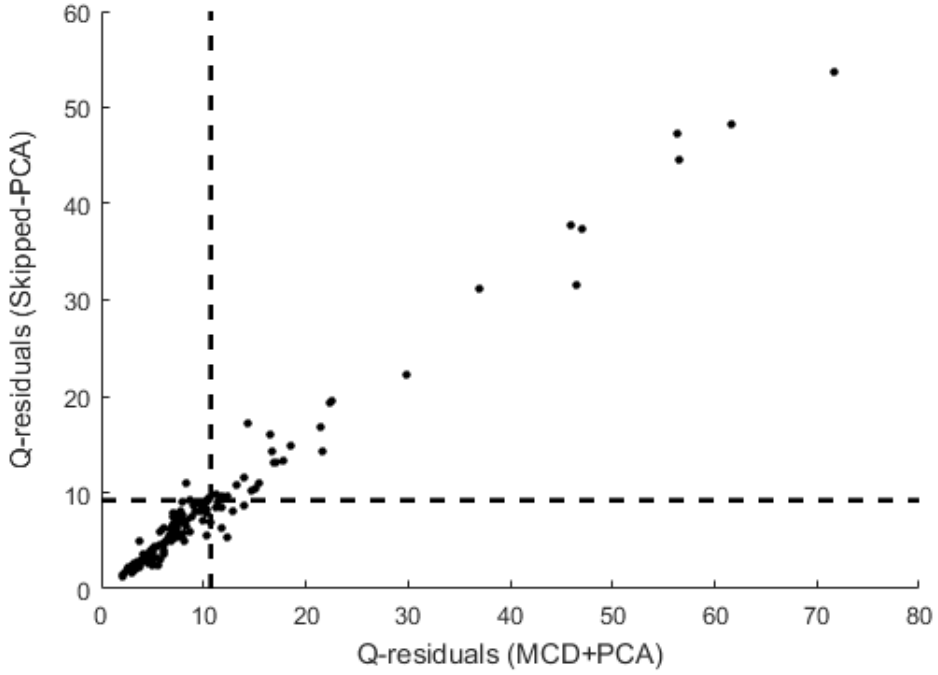

Figure 5.23. Q-residuals of Skipped-PCA method vs. MCD+PCA method testing.

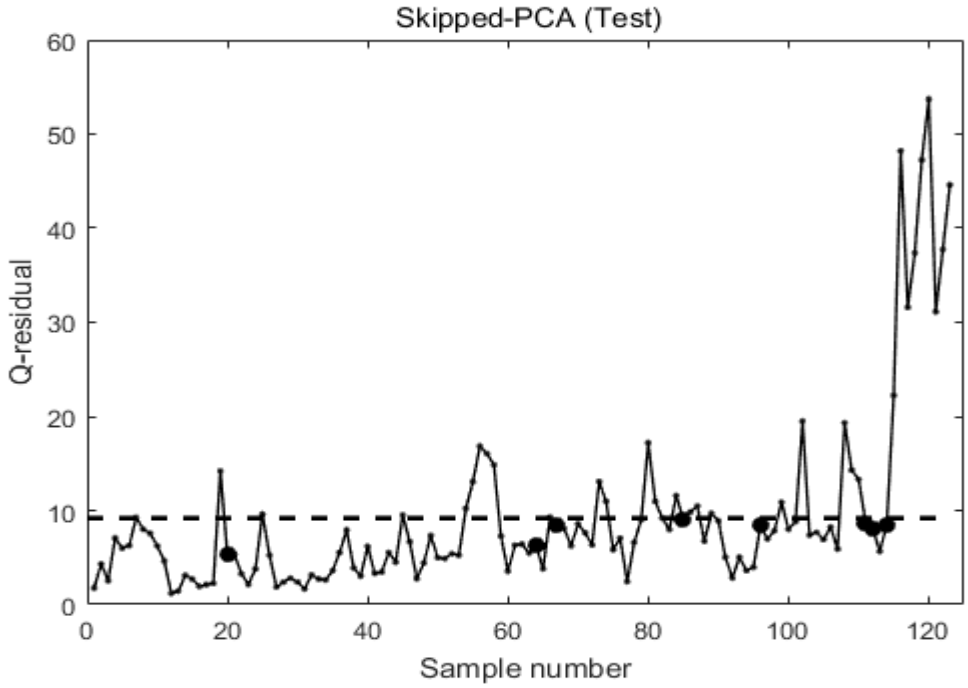

Figure 5.24. Q-residuals of Skipped-PCA method testing.

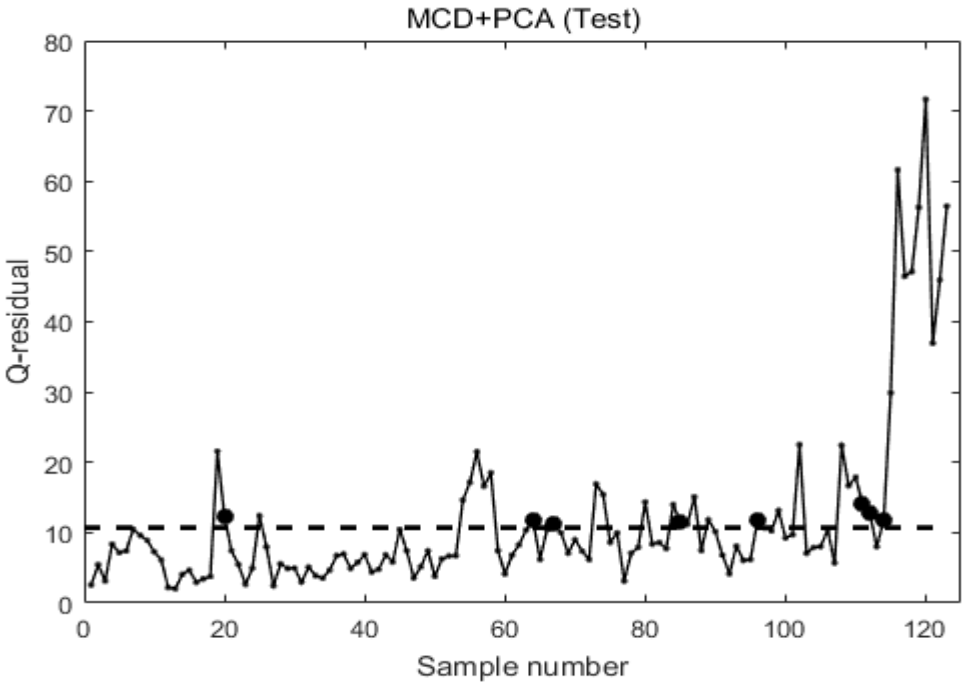

Figure 5.25. Q-residuals of MCD+PCA method testing.

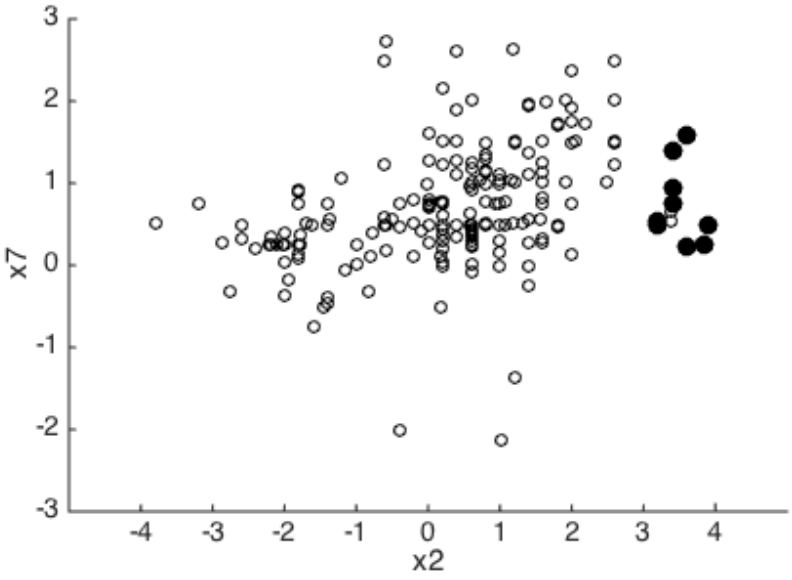

Figure 5.26. x_7 vs x_2 for the test data.

5.2. Monte Carlo Simulations

In this section, predictive performances of LS and various robust regression methods are evaluated using MC simulations. The following regression model is assumed:

$$Y_{ref} = 1 + 2x_1 - x_2 + \varepsilon \tag{5.1}$$

For training and testing the model, 200 and 10^5 (x_1, x_2) data points were randomly selected from two N(0, 1) distributions with a correlation of 0.3. Random error terms are assumed to be N(0, 1) and N(2, 1) distributed for "clean" and contaminated data, respectively. Contaminated data comprise 10% of the whole set. Two versions of the test set are produced: one containing only clean observations, and the other one containing 10% contaminations, as in the training data set. Clean test set is produced to test the accuracy of predictions, while contaminated test set serves the purpose of assessing the metrics, such as root mean squared error (RMSE) and mean absolute error (MAE), used to evaluate the quality of predictions under contamination. MC simulations are repeated for 2000 different datasets, and means of the results are reported.

For robust regression methods, in the current study, firstly, LS model was constructed. Secondly, LTS models were constructed with 10%, 20%, 30%, 40%, 50% trimming with 97.5% CL. The detected outliers by LTS models were removed from the dataset by various methods, and LS method was applied to remaining data. This method is called as "LTS+LS". Finally, reweighted LTS and LTS with a modified two-step procedure were applied. In these methods, constructed LTS model was used as basis. Designations of models are given in Table 5.3.

In order to select the best method for prediction, test results of all models with clean test data were compared in Figure 5.27. It is seen that none of the pure LTS estimators and modified-two-step LTS estimators give smaller RMSE values compared to LS model. On the other hand, LTS10%+LS, LTS20%+LS, and all the reweighted LTS estimators yield smaller predictions compared to those from LS model. One may say that prediction quality gets worse as trimming percentage in LTS models reach 40-50%, hence it's not recommended to use LTS models with more than 30% trimming. It should also be noted

that as contamination percentage in the data is increased, the difference in the prediction performance between LTS-based estimators and LS estimators will increase in favor of LTS based estimators.

Table 5.3. Predictive models used in MC simulations.

Model Notation	**Model Definition**
L_c	LS
Lt1_c	LTS10%
Lt2_c	LTS20%
Lt3_c	LTS30%
Lt4_c	LTS40%
Lt5_c	LTS 50%
LtL1_c	LTS10% + LS
LtL2_c	LTS20% + LS
LtL3_c	LTS30% + LS
LtL4_c	LTS40% + LS
LtL5_c	LTS 50% + LS
LtL1_nc	Reweighted LTS10%
LtL2_nc	Reweighted LTS20%
LtL3_nc	Reweighted LTS30%
LtL4_nc	Reweighted LTS40%
LtL5_nc	Reweighted LTS 50%
LtL1_n2c	Modified two step LTS10%
LtL2_n2c	Modified two step LTS20%
LtL3_n2c	Modified two step LTS30%
LtL4_n2c	Modified two step LTS40%
LtL5_n2c	Modified two step LTS 50%

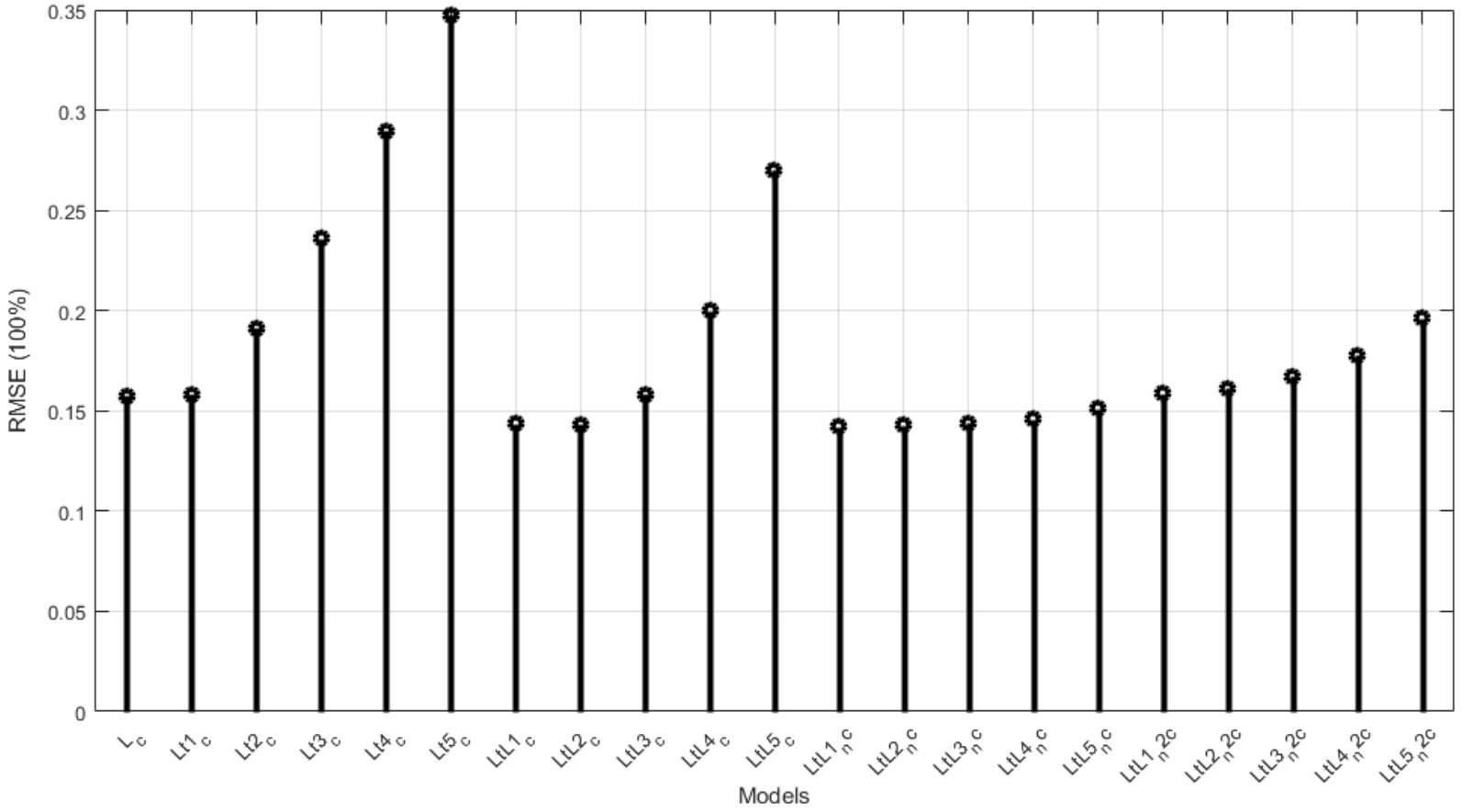

Figure 5.27. RMSE values of various models tested with clean data.

In order to compare the reliability of prediction measures, i.e. of applied methods, mean absolute error (MAE) and root mean squared error (RMSE), in assessing the prediction quality of models, MAE and RMSE values were computed for different percentiles of ranked prediction errors in values test sets including contaminations. MAE and RMSE values of different residual percentages were compared. For this purpose, absolute values of residuals prediction errors were written sorted in ascending order and MAE and RMSE values of the first 100%, 90%, 80%, 70%, 60% and 50% of residuals errors were used to calculate MAE and RMSE computed. The dataset used in simulation study was created with random variables so whenever the program runs, dataset changes. For this reason, by running the MATLAB code for 2000 times, average MAE and RMSE values of runs were calculated and used to compare models.

Figure 5.28 shows the RMSE values of various models for different percentages of ranked absolute residuals. In Figure 5.27, LTS20%+LS was found to be the best predictive model, but RMSE values using all of the ranked absolute prediction errors shows that the predictive performances of LS, LTS10%+LS, LTS20%+LS and LTS30%+LS are indistinguishable. It is important to note that LS model predictions are at least as accurate as those of LTS10%+LS, LTS20%+LS, when all of the RMSE of all of the prediction errors are computed. However, when RMSE values of the first 90% of the ranked prediction errors are taken into consideration, LTS10%+LS and LTS20%+LS methods are seen to have smaller RMSE values, consistent with the results obtained for clean test data. As the percentage of error terms in RMSE is decreased to 50%, LTS30%+LS model is seen to have a comparable RMSE with those of LTS10%+LS and LTS20%+LS, showing that excessively decreasing the percentage of error terms in RMSE calculation gives erroneous results. This shows that RMSE of 70%-90% the highest absolute prediction errors is a better model assessing method compared to taking RMSE of all the prediction error terms, in order to evaluate the predictive performance of models on contaminated test data. In Figure 5.29, similar computations are performed using MAE. It is seen that results obtained by MAEs of all the prediction errors are compatible with the results obtained clean test data: LTS10%+LS, LTS20%+LS have clearly better prediction performance compared to LS. As a smaller percentage of prediction errors are taken into consideration for MAEs, the same picture is consistently obtained, particularly for 70% to 90% of the

residuals range. This shows that using MAE as a measure of predictive performance on test data is an alternative to using RMSE on all error terms.

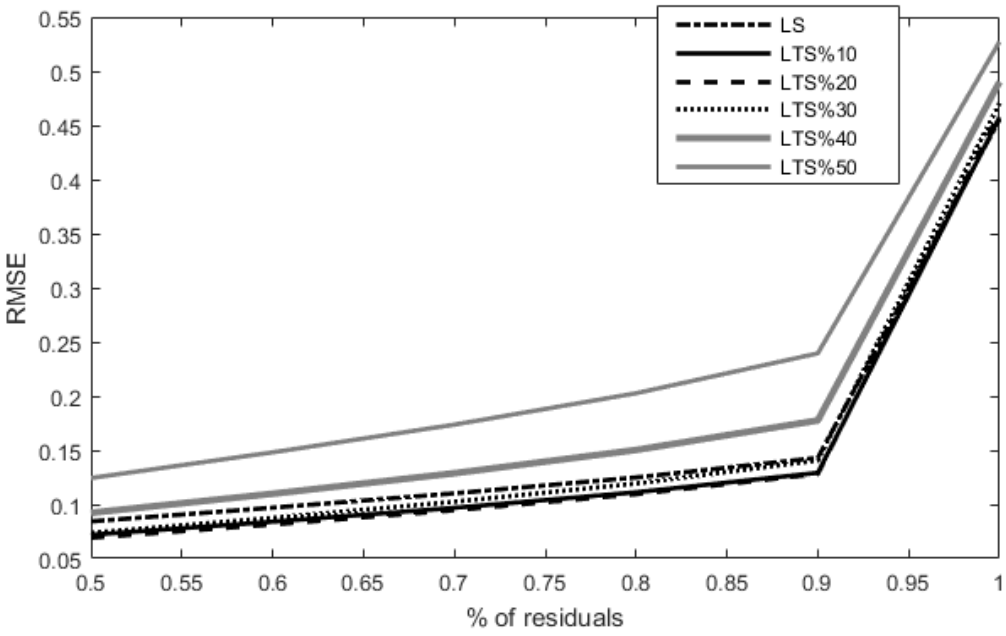

Figure 5.28. RMSE values of LS and LTS+LS models.

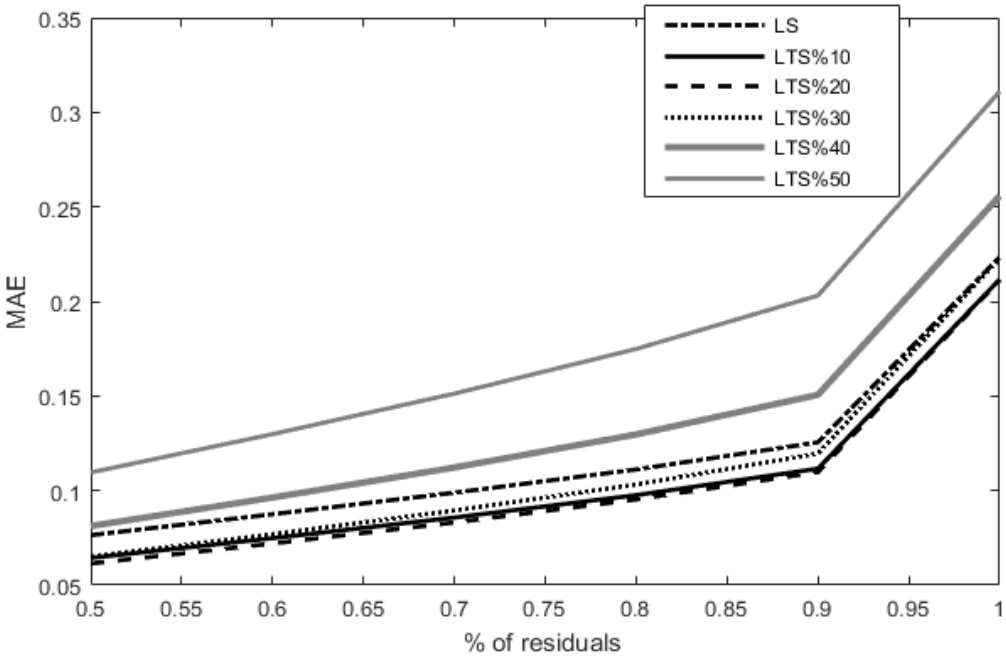

Figure 5.29. MAE values of LS and LTS+LS models.

5.3. Model Prediction

In this section, it is aimed to estimate HD T95 values using the process variables (Table 4.1 and Figure 3.3) measured in CDU. First, conventional LS method, which is known to be sensitive to outliers, (see Section 2.1) was used. Then, robust regression methods, LTS, GM-regression and S-regression, were employed on the same set of data, and predictive performance of the conventional and robust methods was compared. In constructing the predictive models, as employed in the previous sections, the first 200 data points in the historical dataset were used for training the models, while the last 123 data points were used for testing the models.

5.3.1. Modeling Using LS

Here, process variables used as predictors in the model were selected with the help of stepwise regression. First order, interaction and quadratic terms of process variables were included into the model, using the p-values in stepwise regression. Furthermore, R^2 and RMSE values of the model were taken into consideration. The resulting model consisted of nine process variables, and 15 predictors, including three cross product and three quadratic terms. Point estimates, standard error of coefficients (SE) and p-values of the LS model parameters are shown in Table 5.4, Table 5.5 and Table 5.6 respectively. R^2 statistics of the model is 0.49, with MAE and RMSE statistics equal to 3.67 and 4.62 ^{o}C. The process variables used in LS model are x_2 (desalter pressure), x_3 (2nd group heat exchangers exit temperature), x_9 (flow to Naphtha Splitter column), x_{16} (Fired heater transfer temperature 2), x_{18} (Kerosene temperature), x_{19} (LD temperature), x_{21} (LD flow), x_{22} (HD flow) and x_{23} (Top reflux flow). The resulting model is plausible in terms of expected causal relations. As seen in Table 5.4, while there is positive correlation between flow to Naphtha splitter column (x_9) and HD T95 value, showing that HD T95 increases with the upward shift of products in the column yielding heavier products. LD temperature (x_{19}) is also positively correlated with HD T95. LD is drawn from upper tray of the HD, so increase in temperature of LD would necessarily increase the HD temperature and HD T95 value. Increase in top reflux flow (x_{23}) may render lighter components to stay in the bottom of the column, so there may be downward movement of products inside the column. This is

likely to yield a lower HD T95 value, consistent with the sign of estimated model parameter.

Table 5.4. Estimates of LS model parameters.

Model parameters	**Estimate**	**Model parameters**	**Estimate**
Intercept	59690	x_{22}	0.7347
x_2	26.719	x_{23}	-0.0217
x_3	6.7078	$x_2 \times x_3$	-0.2189
x_9	0.0229	$x_{16} \times x_{22}$	-0.0018
x_{16}	-353.91	$x_{21} \times x_{22}$	5.908×10^{-5}
x_{18}	-2.2304	x_3^2	-0.0127
x_{19}	2.1702	x_{16}^2	0.52148
x_{21}	-0.1245	x_{22}^2	-3.144×10^{-5}

Table 5.5. SE of LS model parameters.

Model parameters	**SE**	**Model parameters**	**SE**
Intercept	15665	x_{22}	0.344
x_2	6.879	x_{23}	0.00324
x_3	1.368	$x_2 \times x_3$	0.0587
x_9	0.004	$x_{16} \times x_{22}$	0.00098
x_{16}	91.126	$x_{21} \times x_{22}$	1.93×10^{-5}
x_{18}	0.481	x_3^2	0.004
x_{19}	0.423	x_{16}^2	0.132
x_{21}	0.034	x_{22}^2	9.51×10^{-6}

Table 5.6. P-values of LS model parameters.

Model parameters	**p-value**	**Model parameters**	**p-value**
Intercept	0.00019	x_{22}	0.03422
x_2	0.00014	x_{23}	2.698×10^{-10}

Model parameters	p-value	Model parameters	p-value
x_3	2.075×10^{-6}	$x_2 \times x_3$	0.00025
x_9	3.774×10^{-8}	$x_{16} \times x_{22}$	0.06065
x_{16}	0.00014	$x_{21} \times x_{22}$	0.00252
x_{18}	6.656×10^{-6}	x_3^2	0.00177
x_{19}	7.361×10^{-7}	x_{16}^2	0.00012
x_{21}	0.00034	x_{22}^2	0.00114

Experimental and fitted HD T95 values using the LS model are shown in Figure 5.30. Though the LS fit generally, captures the time trend of the T95 values, residuals of the last ~20 observations seem to be high, and there seems to occasional spikes in HD T95, which have not been captured well.

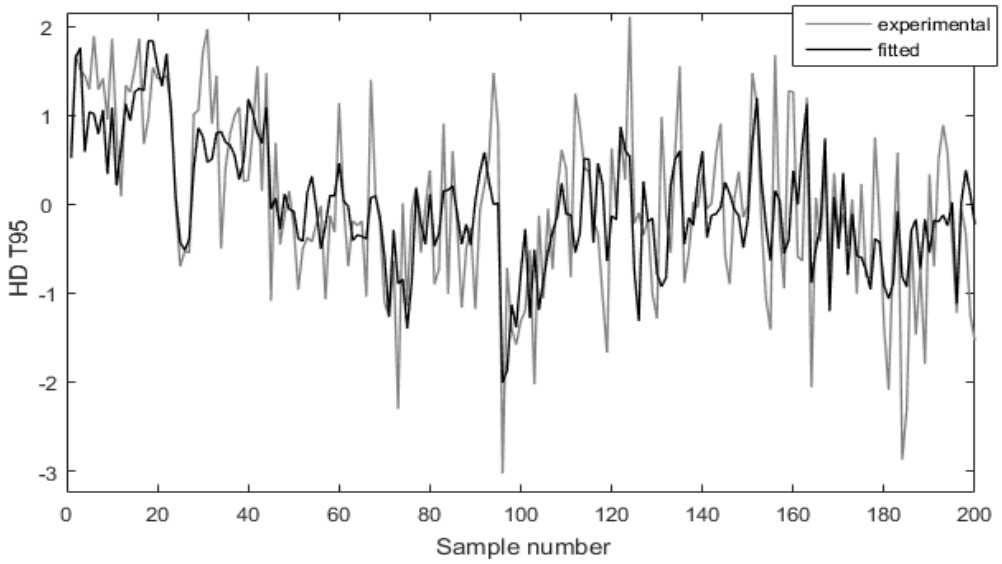

Figure 5.30. Experimental and fitted HD T95 values of training set by LS method.

Figure 5.31 shows the scatter plot of the experimental vs. fitted T95 values (filled circles to be explained in the next subsection). It is desired that T95 values fall close to the 45^o dashed line. Though one cannot observe a trend in the residuals, the wide scatter of the points brings doubt on the fitting quality of the model.

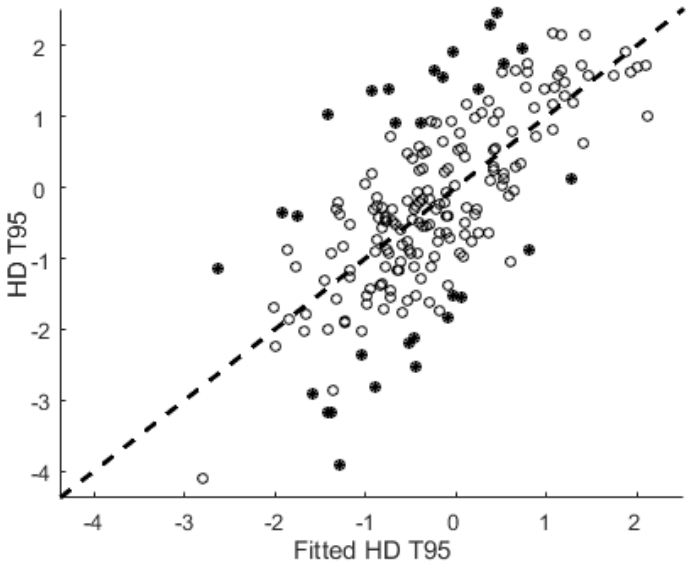

Figure 5.31. Experimental vs. fitted HD T95 values of LS model on training set.

Normal probability plot of residuals (Figure 5.32a) is used to check whether model error terms are normally distributed. Though one is temped to conclude that residuals are non-normally distributed, since the only the middle portion of the distribution fits well to the line representing normal distribution, normal probability plot of a simulated random sample of normal distributed data consisting of 200 observations (Figure 5.32b) shows that a certain degree of discrepancy between normal distribution and small sample distribution from a normal distributed population is unavoidable. Hence, one cannot conclude that the residuals are not normally distributed.

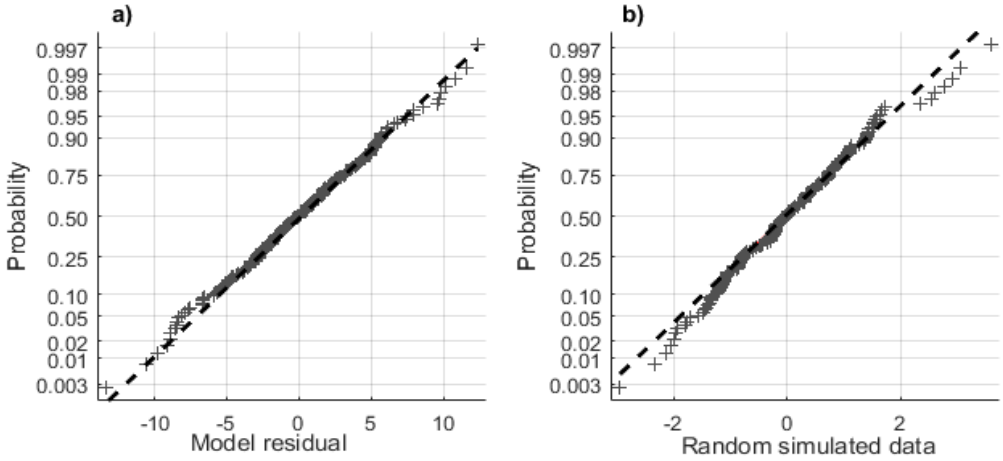

Figure 5.32. (a) normal probability plot of LS model residuals and (b) simulated random data of 200 points.

Figure 5.33 shows the standardized LS residuals vs. MD. Here, it is aimed to examine potential outliers, both in x- and y-space, influencing the LS model. Here the vertical and horizontal dashed lines represent 97.5% confidence limits for residuals, found to be equal to 2.26 using Student's t-distribution, and MD, found to be equal to 4.36 using the square root of a chi-square distribution with 9 degrees of freedom. Data points on the upper left (with respect to dashed lines) are vertical outliers. Data points on the lower right are "good" leverage points, while those on the upper right are "bad" leverage points. Observations exceeding MD limit, 96, 136, 160, 161, 162, and 168 are considered to be "good" leverage points, since none of these points exceeds the regression residual limit. Overall, there are 4 observations, which comprise 2% of all data, out of the CLs. However, it should be recalled that LS residuals and MD may suffer from masking effect, so one cannot be confident that data points within the CLs are free from contamination [18].

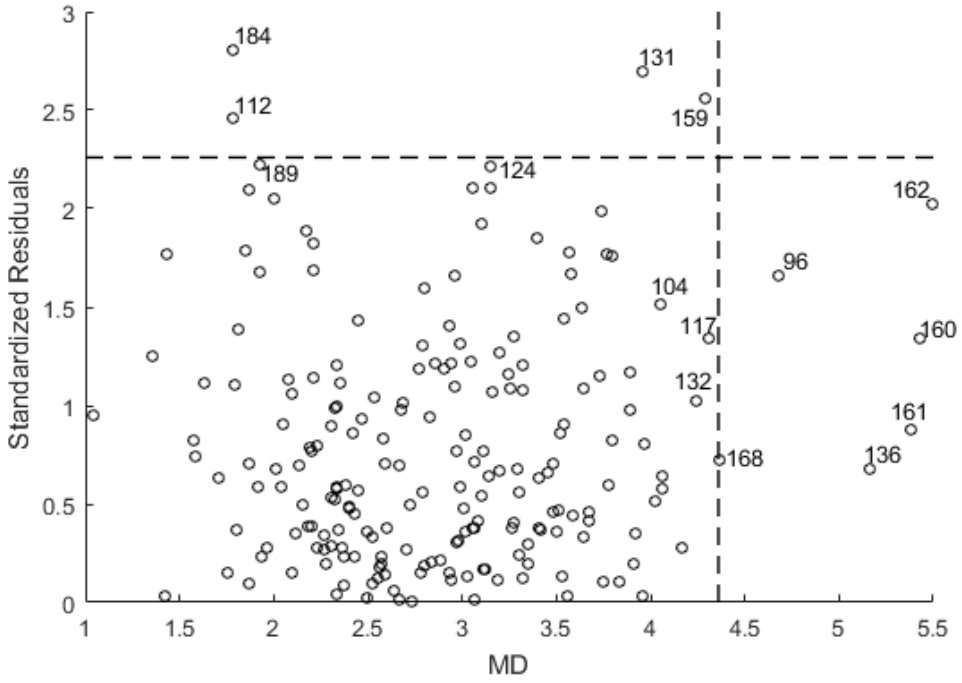

Figure 5.33. Plot of LS residual versus Mahalanobis distance.

Predictive performance of the constructed LS model was evaluated using the test data. When the LS model was applied to the test data, MAE and RMSE statistics were found to be equal to 3.93 and 5.09 ^{o}C. It should be recalled that MAE and RMSE statistics for the training set were found to be equal to 3.67 and 4.62 ^{o}C respectively, hence

prediction errors are slightly higher than the fitted model errors, leaving room for predictive model improvement.

Experimental and predicted T95 trajectories of the test data are shown in Figure 5.34 and compared in Figure 5.35. Superposition of test and predicted HD T95 values (Figure 5.34) shows the existence of poor predictions during a number of time intervals.

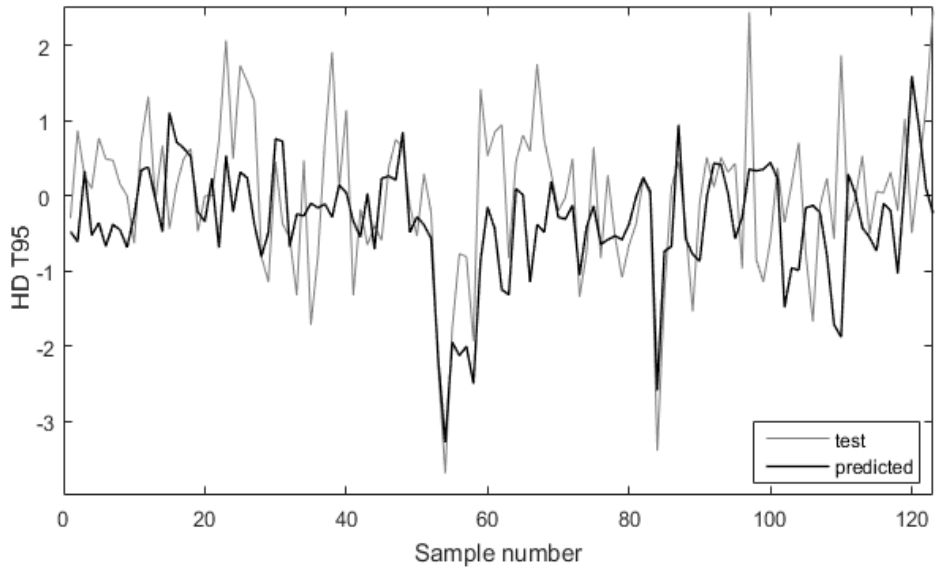

Figure 5.34. Experimental and predicted HD T95 values of LS model for the test data.

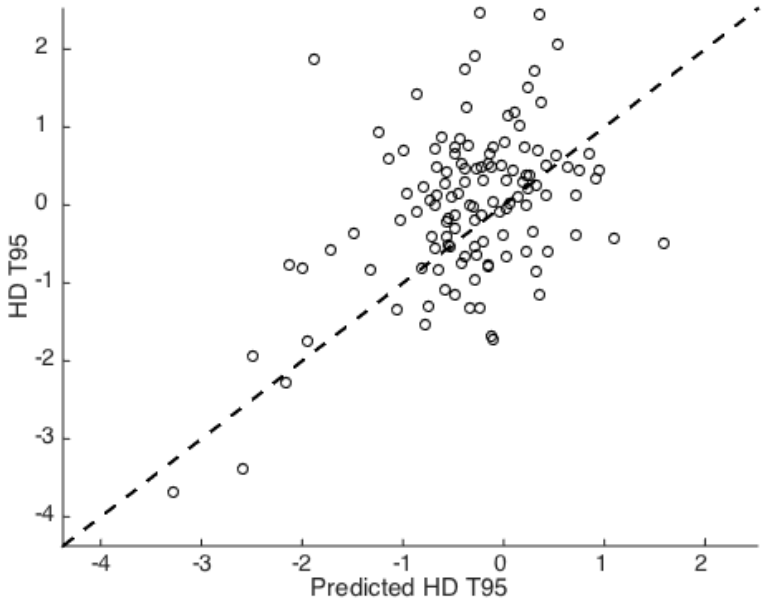

Figure 5.35. Test and predicted HD T95 values of LS model.

Standardized LS prediction residuals vs. MD statistics for the test set are shown in Figure 5.36. It is seen that 110th and 123rd data points have high MD statistics, meaning that these points are far from the bulk of the training data in the predictor space, and high prediction residuals, showing that experimental and predicted T95 values of these points highly differ. Moreover, data points 38, 59, 62, 67 are on the upper left region, and hence can be regarded as outliers in the test set. Furthermore, it is seen that these six poor predicted samples in the test dataset are all underpredicted, showing that there may be a bias in the predicted values.

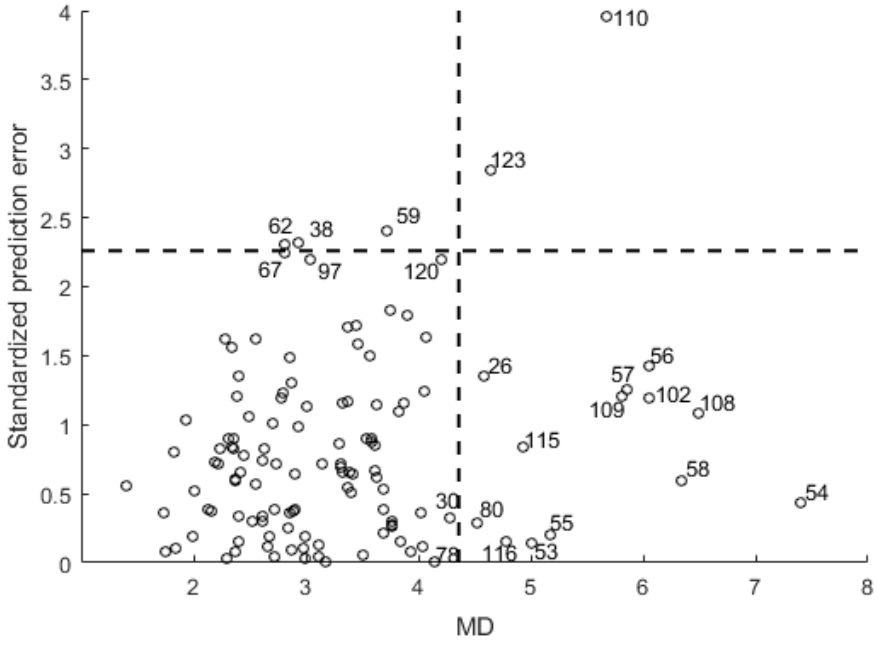

Figure 5.36. Plot of LS prediction error versus Mahalanobis distance.

5.3.2. Modeling Using Robust Regression

Here, robust regression methods, LTS, GM-regression and S-regression, were applied and their predictive performance was compared with that of LS model. On MATLAB, FDSA toolbox was used for LTS and S-regression, while Statistics and Machine Learning Toolbox was used for GM regression [47]. In all the reported LTS models, outliers are first trimmed at the given confidence level, and then LS models are fit to the remaining data. To assess the predictive quality of the models, MAE and RMSE values at different percentages of ranked absolute residuals were computed. RMSE and

MAE of the model residuals and prediction errors obtained from various LS and robust regression models are shown in Table 5.7, and Table 5.8, where, the regressions denoted LTS10%, LTS20%, LTS30% and LTS40% are actually LTS10%+LS, LTS20%+LS, LTS30%+LS and LTS40%+LS.

In the LTS models, trimming parameter h was selected as 10%, 20%, 30% and %40. Outliers in each reweighted LTS method was identified using confidence limits of 95%, 97.5% and 99%. The initial model, used by the robust methods, was obtained by the LS method (see the previous section). Since LTS models were obtained using random search, a large number of initial seeds, as much as 10^5, were used to obtain reliable results. GM-estimator was applied for three different tuning values of Huber estimation: 1.1, 1.345, 1.5 while S-estimator was applied at 97.5% confidence level with three different breakdown point: 20%, 30%, 40%.

When the RMSE and MAE values of the model residuals are compared, LTS20% at 95% CL and LTS30%at 97.5% CL models have slightly lower MAE 100% values compared to that of LS model, which has the smallest RMSE 100% values (Table 5.7). It should be recalled that LS solution guarantees that RMSE 100% values is minimized among all possible linear estimators. As lower percentages of error terms considered, for instance at RMSE 80%, LS model gives a RMSE value of 3.01 ^{o}C, while LTS20% at 95% CL, LTS30% at 97.5% CL and S-estimator at BDP of 40% give RMSE values of 2.86, 2.84 and 2.83 ^{o}C, respectively. These results suggest that RMSE at lower percentages than 100% and MAE of residuals should also be checked to evaluate the quality of fitting of the models to the data.

As seen in Table 5.8, most of the robust regression methods yield prediction errors with lower MAE and RMSE values when compared with LS method. Even using RMSE 100% values, LTS20% at 95% CL, LTS30% at 97.5% CL models are seen to make superior predictions compared to LS model. When RMSE 80% values, for instance, are compared, LS model gives a predictive RMSE of 3.18 ^{o}C, while LTS20% at 95% CL, and LTS30% at 97.5% CL models give predictive RMSEs of 2.96 and 2.96 ^{o}C. A similar picture emerges, when MAE values are compared: using 100% of test data, MAE values of LS, LTS20% at 95% CL, and LTS30% at 97.5% CL models are 3.93, 3.74 and 3.75 ^{o}C,

respectively. It should also be noted that S-estimator at BDP of 40%, which have yielded promising results during fitting procedure, was only marginally superior to LS model. As a result, LTS30% with 97.5% CL was selected as the best model for predicting HD T95. The following analysis focuses solely on this estimator.

Removing the outliers by the LTS30% model at 97.5% CL, and employing LS analysis on the remaining data yields a model with a R^2 of 0.69. The estimates, SE of model parameters and their P-values are listed in Table 5.9, Table 5.10 and Table 5.11, respectively. When compared with LS method, none of the parameter estimates have changed signs, and most of the terms have increased their significance, while only x_{22} and x2×x3 terms have decreased their significance. When SE values are compared, it is seen that almost all SE values are less than LS model. The SE values were taken into consideration when comparing and examining the estimators of LS and LTS30%. For instance, parameter estimate of x_2 is 26.7 and 19.3 (Table 5.4) in LS and LTS30% models, respectively. SEs of parameter estimate of x_2 are found to be 6.8 and 5.1 in LS and LTS30%, respectively, showing that SE has not significantly changed, but the parameter estimate has changed by more than one SE. Furthermore, parameter estimates of x_9, x_{22}, $x_2 \times x_3$, $x_{16} \times x_{22}$, and x_3^2 have also changed by more than one SE upon robust regression. These show that robust regression has significantly changed the model parameter estimates.

Figure 5.37 shows the trajectories of experimental and fitted HD T95 values using the LTS30% model. Fitted values, generally, captures the time trend of the experimental HD T95 values. 30 outliers are detected by LTS30% model, comprising ~15% of the training data, and these outliers are indicated with filled circles in Figure 5.37. Most of the outlier data seem to be sampled from the highest and lowest HD measurements, as represented by spikes with respect to time. Moreover, more than ten of the outliers are from successive samples of two observations. This hints the possibility of a dynamic effect of contamination, i.e. the experimental conditions of the current daily measurement may be correlated with the experimental conditions of the next daily measurement.

Table 5.7. RMSE and MAE of the model residuals by various robust estimators.

	LS Model	LTS10%			LTS20%			LTS30%			LTS40%			M Estimator (Huber)			S Estimator (CL .975)		
		CL 95%	CL 975%	CL 99%	CL 95%	CL 975%	CL 99%	CL 95%	CL 975%	CL 99%	CL 95%	CL 975%	CL 99%	TV 1.1	TV 1.345	TV 1.5	bdp.20	bdp.30	bdp.40
MAE 100%	3.67	3.66	3.64	3.65	3.62	3.64	3.66	3.62	3.62	3.66	4.17	3.70	3.65	3.62	3.64	3.64	3.61	3.60	3.62
MAE 90%	3.00	2.94	2.95	2.97	2.84	2.91	2.94	2.85	2.84	2.91	3.24	2.92	2.87	2.93	2.96	2.97	2.89	2.84	2.84
MAE 80%	2.54	2.48	2.48	2.50	2.34	2.42	2.46	2.29	2.32	2.41	2.54	2.33	2.30	2.44	2.48	2.49	2.41	2.32	2.27
MAE 70%	2.14	2.10	2.08	2.11	1.94	2.03	2.07	1.87	1.91	2.01	1.92	1.85	1.87	2.06	2.09	2.10	2.01	1.93	1.83
MAE 60%	1.78	1.73	1.71	1.74	1.58	1.67	1.71	1.50	1.55	1.67	1.44	1.47	1.45	1.69	1.72	1.73	1.64	1.56	1.46
MAE%50	1.46	1.40	1.39	1.43	1.24	1.34	1.39	1.19	1.21	1.36	1.13	1.18	1.09	1.36	1.40	1.41	1.31	1.22	1.15
RMSE 100%	4.62	4.71	4.65	4.64	4.79	4.73	4.73	4.82	4.79	4.76	5.73	4.95	4.91	4.63	4.62	4.62	4.67	4.76	4.86
RMSE 90%	3.62	3.57	3.60	3.61	3.55	3.56	3.58	3.63	3.58	3.57	4.32	3.76	3.71	3.58	3.60	3.60	3.56	3.56	3.64
RMSE 80%	3.01	2.95	2.96	2.98	2.86	2.89	2.93	2.82	2.84	2.89	3.33	2.92	2.89	2.92	2.96	2.97	2.90	2.83	2.83
RMSE 70%	2.52	2.48	2.47	2.48	2.35	2.41	2.45	2.28	2.33	2.37	2.43	2.25	2.35	2.44	2.48	2.49	2.41	2.34	2.25
RMSE 60%	2.06	2.03	2.00	2.01	1.91	1.96	1.99	1.80	1.87	1.95	1.72	1.73	1.81	1.98	2.01	2.02	1.94	1.87	1.77
RMSE 50%	1.67	1.63	1.62	1.64	1.4854	1.53	1.60	1.42	1.46	1.57	1.29	1.37	1.33	1.57	1.62	1.63	1.52	1.45	1.39

Table 5.8. RMSE and MAE of the prediction errors by different estimators.

	LS Model	LTS10%			LTS20%			LTS30%			LTS40%			M Estimator (Huber)			S Estimator (CL .975)		
		CL 95%	CL 975%	CL 99%	CL 95%	CL 975%	CL 99%	CL 95%	CL 975%	CL 99%	CL 95%	CL 975%	CL 99%	TV 1.1	TV 1.345	TV 1.5	bdp.20	bdp.30	bdp.40
MAE 100%	3.93	3.84	3.96	3.96	3.74	3.74	3.93	3.95	3.75	4.12	4.22	4.15	4.26	3.86	3.89	3.91	3.88	3.81	3.93
MAE 90%	3.16	3.07	3.18	3.19	3.02	2.98	3.16	3.17	3.02	3.30	3.54	3.38	3.41	3.11	3.14	3.15	3.11	3.05	3.15
MAE 80%	2.65	2.57	2.69	2.68	2.52	2.49	2.64	2.63	2.50	2.76	3.04	2.84	2.84	2.61	2.63	2.64	2.62	2.56	2.62
MAE 70%	2.24	2.19	2.25	2.25	2.16	2.10	2.22	2.22	2.13	2.33	2.59	2.46	2.43	2.21	2.23	2.23	2.21	2.17	2.22
MAE 60%	1.88	1.83	1.89	1.88	1.82	1.75	1.85	1.84	1.79	1.95	2.13	2.12	2.05	1.86	1.87	1.87	1.85	1.80	1.85
MAE%50	1.53	1.50	1.54	1.53	1.51	1.40	1.51	1.47	1.49	1.60	1.71	1.79	1.69	1.52	1.53	1.53	1.51	1.44	1.47
RMSE 100%	5.09	4.99	5.13	5.14	4.81	5.09	5.12	5.16	4.84	5.36	5.23	5.21	5.55	5.00	5.04	5.06	5.03	4.97	5.12
RMSE 90%	3.84	3.70	3.85	3.87	3.64	3.82	3.84	3.88	3.67	4.01	4.26	4.01	4.12	3.77	3.80	3.82	3.77	3.72	3.85
RMSE 80%	3.18	3.06	3.22	3.22	2.96	3.15	3.16	3.16	2.97	3.31	3.67	3.28	3.35	3.12	3.15	3.16	3.14	3.07	3.14
RMSE 70%	2.67	2.58	2.67	2.68	2.52	2.65	2.63	2.65	2.50	2.76	3.14	2.81	2.85	2.63	2.65	2.66	2.63	2.60	2.66
RMSE 60%	2.24	2.14	2.23	2.23	2.11	2.18	2.18	2.20	2.09	2.31	2.57	2.41	2.40	2.22	2.23	2.23	2.20	2.16	2.22
RMSE 50%	1.83	1.75	1.81	1.82	1.74	1.77	1.77	1.74	1.75	1.89	2.07	2.03	1.97	1.83	1.84	1.84	1.80	1.73	1.75

Table 5.9. Estimates of LTS30% model parameters.

Model parameters	**Estimate**	**Model parameters**	**Estimate**
Intercept	65095	x_{22}	0.3852
x_2	19.298	x_{23}	-0.0207
x_3	6.8454	$x_2 \times x_3$	-0.1546
x_9	0.0179	$x_{16} \times x_{22}$	-0.0007
x_{16}	-382.98	$x_{21} \times x_{22}$	6.833×10^{-5}
x_{18}	-1.9565	x_3^2	-0.0172
x_{19}	1.8328	x_{16}^2	0.5607
x_{21}	-0.1363	x_{22}^2	-3.797×10^{-5}

Table 5.10. SE of LTS30% model parameters.

Model parameters	**SE**	**Model parameters**	**SE**
Intercept	12004	x_{22}	0.242
x_2	5.141	x_{23}	0.002
x_3	1.043	$x_2 \times x_3$	0.044
x_9	0.0031	$x_{16} \times x_{22}$	0.0007
x_{16}	69.82	$x_{21} \times x_{22}$	1.49×10^{-5}
x_{18}	0.383	x_3^2	0.003
x_{19}	0.329	x_{16}^2	0.102
x_{21}	0.026	x_{22}^2	7.18×10^{-6}

Table 5.11. P-values of LTS30% model parameters.

Model parameters	**p-value**	**Model parameters**	**p-value**
Intercept	2.23×10^{-7}	x_{22}	0.114
x_2	0.00024	x_{23}	4.61×10^{-14}
x_3	7.60×10^{-10}	$x_2 \times x_3$	0.00062
x_9	4.06×10^{-8}	$x_{16} \times x_{22}$	0.268
x_{16}	1.66×10^{-7}	$x_{21} \times x_{22}$	9.70×10^{-6}

Model parameters	p-value	Model parameters	p-value
x_{18}	9.81×10^{-7}	x_3^2	6.92×10^{-8}
x_{19}	1.09×10^{-7}	x_{16}^2	1.44×10^{-7}
x_{21}	7.76×10^{-7}	x_{22}^2	4.25×10^{-7}

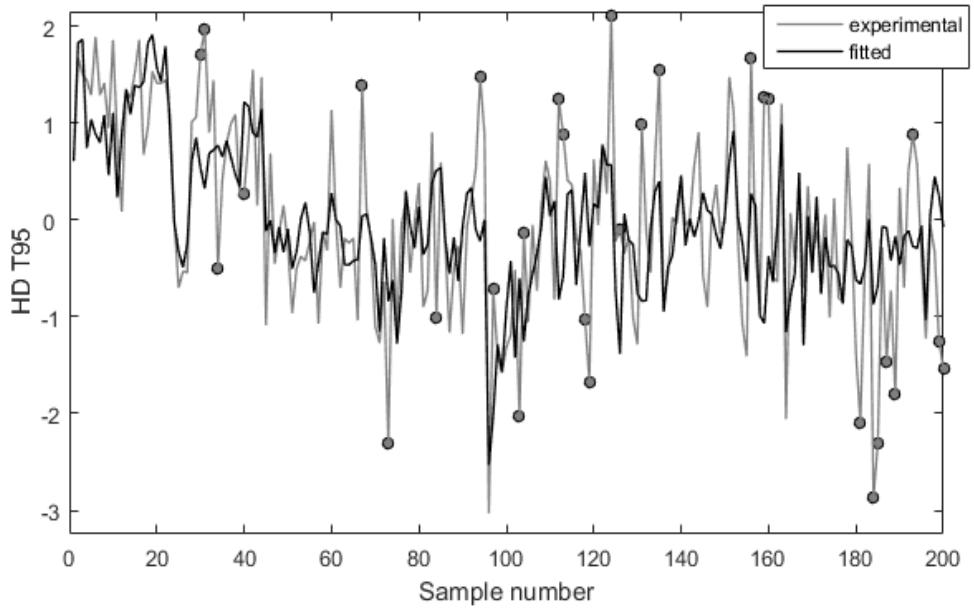

Figure 5.37. Experimental and fitted HD T95 values of training set of LTS30% model.

A MCD model at breakdown point of 25% was constructed to predictor variables x_2, x_3, x_9, x_{16}, x_{18}, x_{19}, x_{21}, x_{22} and x_{23}. MDs (robust distances) were computed using the sample mean and covariance matrix estimates determined by MCD, and robust residuals vs. robust distances were plotted (Figure 5.38). LTS30% outliers were indicated with filled circles and CLs of regression and MDs were drawn at 2.26 and 4.36, respectively. Comparing Figure 5.38 with Figure 5.33 the most significant difference is the location of data points 159 and 160 in the regression residual and MD spaces obtained using conventional and robust statistics. Both of these points are identified as “bad leverage” points and excluded from the regression by LTS30% model, while only sample 159 is barely out of the regression residual CL, and cannot be classified as a leverage point using MD. Data points 124 and 131 also show different locations when Figure 5.38 and Figure 5.33 are compared. While sample 131 is out of the CL in Figure 5.33, it is bad leverage point in Figure 5.38. Similarly, while data point 124 is not out of but very close to CL limit

in Figure 5.33, it is in bad leverage region in Figure 5.38, so samples 124 and 131 are also excluded from the regression by LTS30% model beside data points 159 and 160. Furthermore, there are four more leverage points, which are identified by the LTS30% model, but not by the LS model. These data points seem to be sufficient in perturbing the fitted plane to a significantly different direction, as also evidenced by the changes in some of the parameter estimates (see Table 5.10).

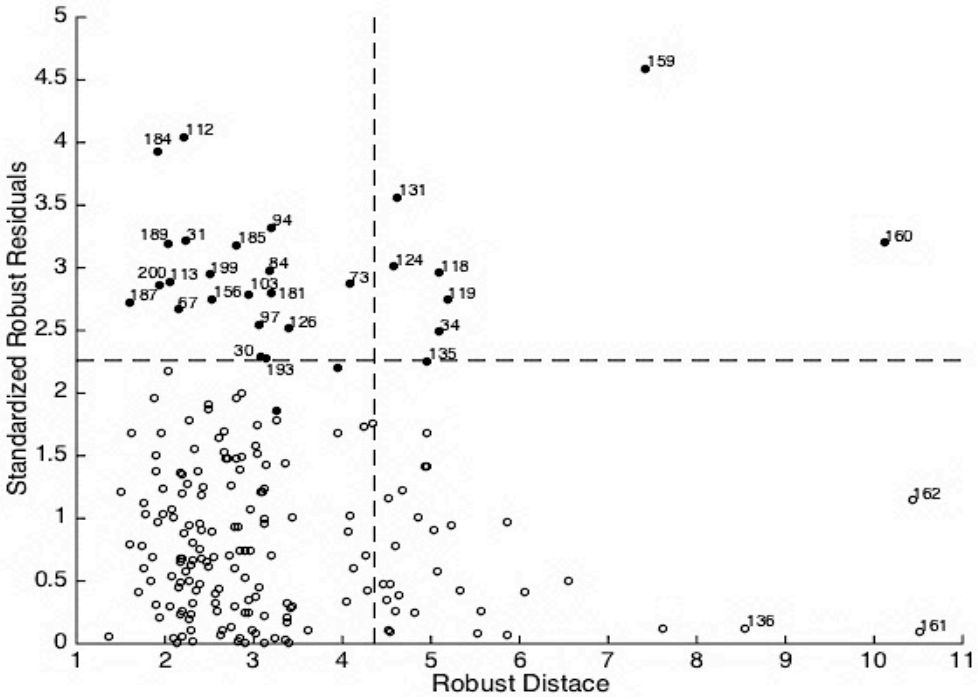

Figure 5.38. LTS30% residuals vs. robust distance.

Similar to Figure 5.31, experimental vs. fitted HD T95 values determined by LTS30% model is plotted in Figure 5.39, in which outliers were indicated with filled circles. It is seen that outliers are generally far to the 45^o dashed line. Figure 5.40 shows the residuals with respect to fitted values. Outlier data points have residuals between 6 ^{0}C and 16 ^{0}C.

For the training set, LTS30% model yielded MAE value as 3.62 ^{0}C. In order to determine robust estimate of RMSE, the relation $RMSE = MAE/0.796$ was used and robust estimate of RMSE was calculated as 4.54 ^{0}C. As stated in Section 3.4, in laboratory tests, reproducibility is an important parameter indicating the difference between two single and independent test results, obtained by different operators working in different

laboratories on identical test material [46]. The relation between reproducibility and RMSE value is given with the equation: $reproducibility = 1.96\sqrt{2}RMSE$. When this equation is used, LTS30% method yields an estimate of 13.0 ^{0}C for the reproducibility, which is slightly higher than the tabulated value (8.48 ^{0}C - 10.9 ^{0}C), but still acceptable.

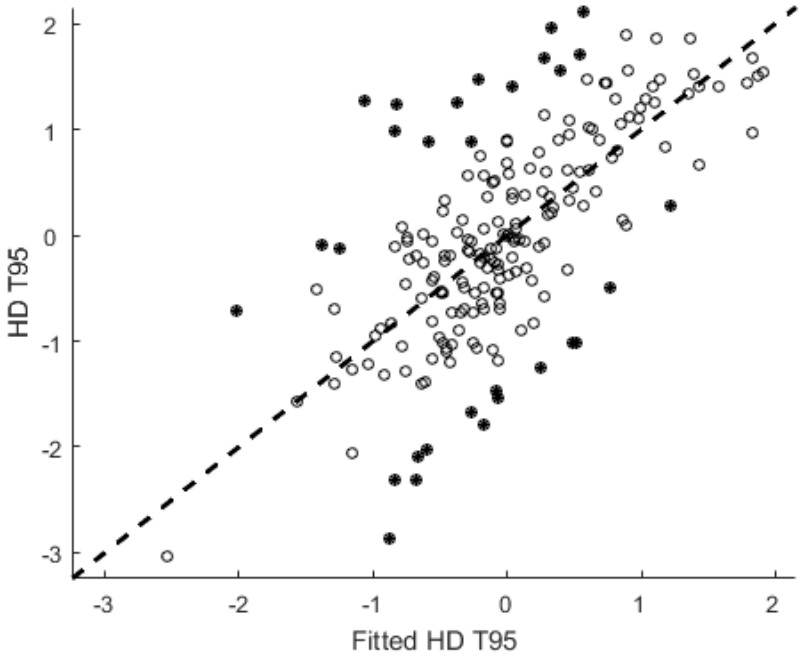

Figure 5.39. Experimental vs. fitted HD T95 values of training set of LTS30% model.

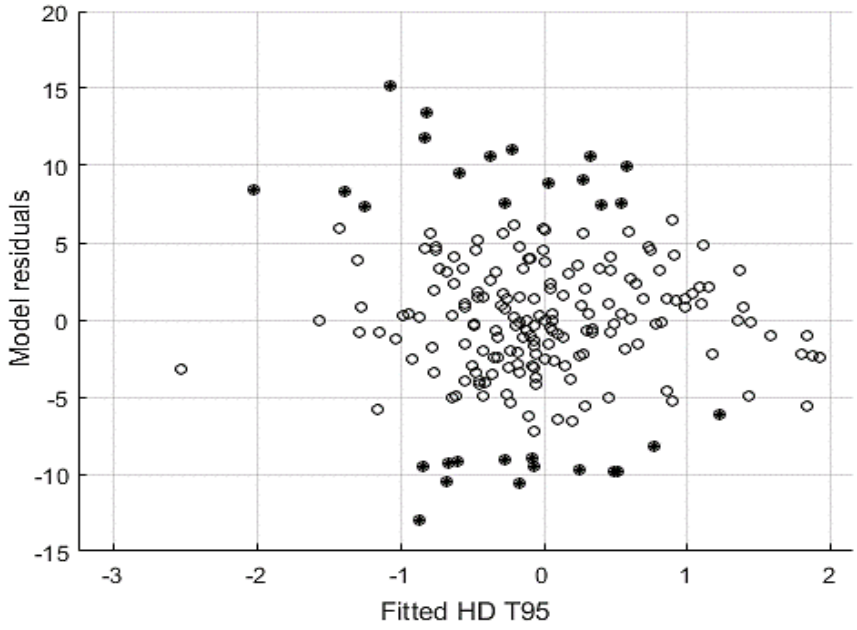

Figure 5.40. Model residuals vs. fitted T95 values of LTS30% Model.

Predictive performance of the constructed LTS30% model was evaluated using test data. When the LTS30% model was applied to the test data, MAE value was found as 3.75 ^{0}C, whereas RMSE value was found as 4.84 ^{0}C, which are slightly higher than the fitted model errors. It is also found that 74% of prediction errors are lower than 5 ^{0}C. Experimental and predicted T95 trajectories of the test data are shown in Figure 5.41 and compared in Figure 5.42. There are a number of poorly predicted samples, similar to that LS model predictions (see Figure 5.34 and Figure 5.35), hence one cannot, visually, discern a striking difference in the predictions of the test set using LS and LTS30% methods.

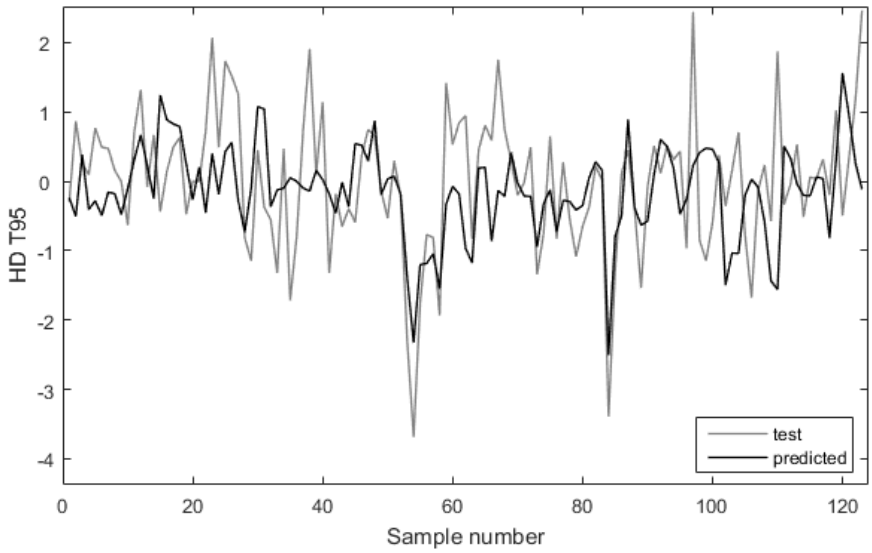

Figure 5.41. Tested and predicted HD T95 values of LTS30% model.

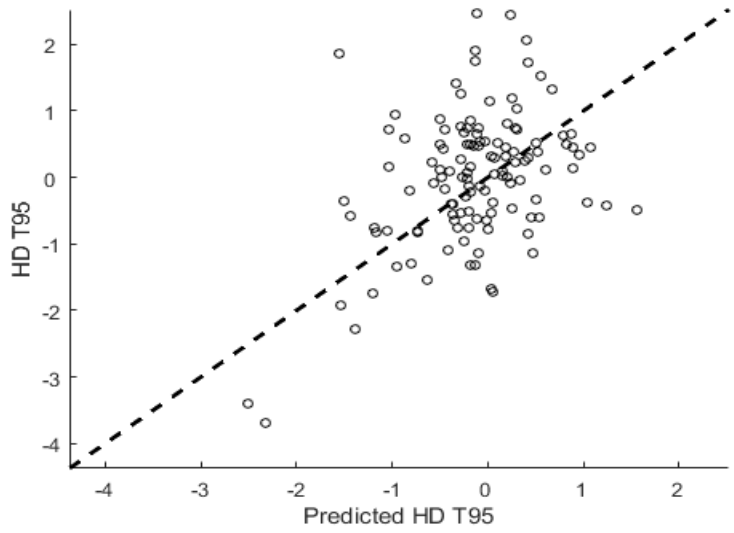

Figure 5.42. Test vs predicted HD T95 values of LTS30% model.

5.3.3. Comparison of fitted and predicted residuals in LS and LTS30% models

Absolute residuals (in the original units of ^{o}C) of the LS and LTS30% model were compared in Figure 5.43. Almost all the absolute residuals above ~7.5 ^{o}C were found to be higher in the LTS30% regression, showing the LTS30% regression moved the LS fitted plane farther away from these data points. While absolute residuals of the rest of the data points are generally in agreement in both regression models, there are four data points, with high LS regression residuals and low LTS30% regression residuals; showing that LTS30% plane is "fine-tuned", as compared to the LS plane, to be more representative of the rest of the data points. Excluding the outliers, LTS30% model can predict HD T95 value with maximum 7 ^{0}C error, which is an efficient estimation when the error range of test method is considered.

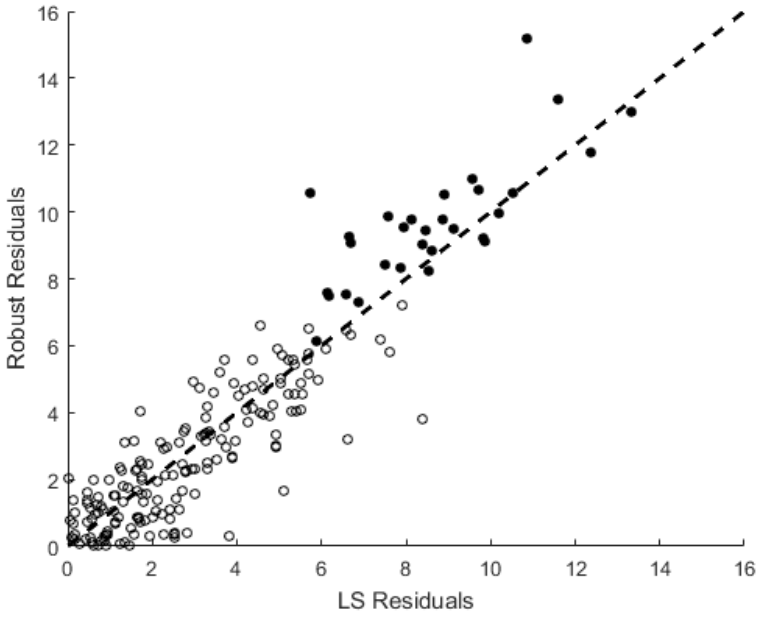

Figure 5.43. Absolute values of LTS30% residuals vs. LS residuals of training data.

LS and LTS30% model predictions of the test data are compared in Figure 5.44. There are two significant differences in the predictions. First, most of test data points, while residing in a narrow band, are consistently and slightly overpredicted by the LTS30% method, compared to LS model. Second, six out seven data points with the lowest T95 predictions by the LS method are significantly overpredicted by the LTS30%.

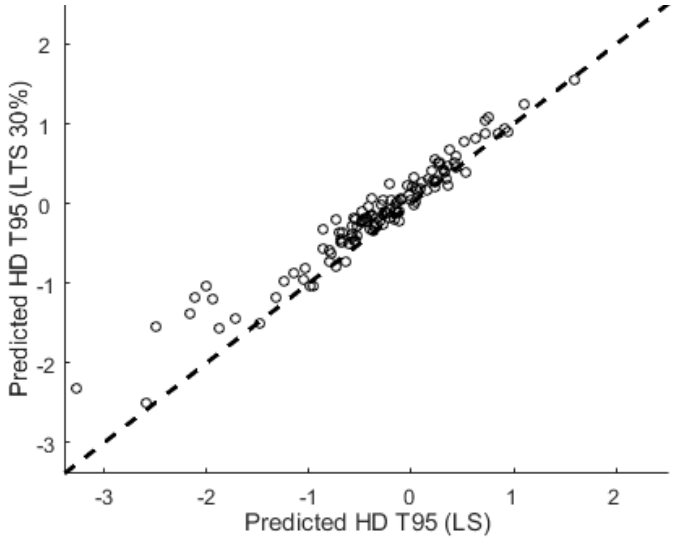

Figure 5.44. LTS30% predicted T95 values vs. LS predicted T95 values of test data.

LS and LTS30% prediction errors of test data were compared in Figure 5.45. Prediction errors from LTS30% model generally are positively biased compared to those from LS model, as also observed in Figure 5.44. Out of the six data points with significant changes in their predictions by the two methods, while the rest of them can be better predicted by the LS method. Medians of the prediction errors of test data were found as 1.46 ^{0}C for LS and 0.04 ^{0}C for LTS30% model. This shows that prediction errors from LTS30% model are practically unbiased, while LS model yields biased predictions, possibly due to contamination bias in the model construction step.

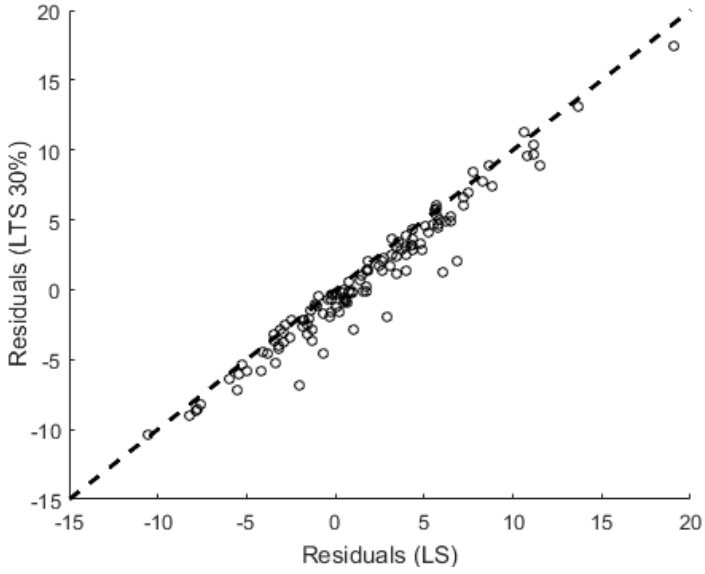

Figure 5.45. LTS30% residuals vs. LS residuals of test data.

6. CONCLUSIONS AND RECOMMENDATIONS

In this thesis, conventional and robust statistical methods are used for monitoring and HD T95 prediction of a CDU process in TUPRAS İzmit Refinery. Trajectories of the process variables are obtained from TUPRAS historical database for a one-year period, and on-line process variable measurements are averaged over 4 hours about the laboratory sampling times. Out of the totally collected 323 observations, the first 200 observations were used for constructing the exploratory and predictive models, while the remaining 123 data points were used for testing the models. Though it is not possible to say anything definite, operational changes, disturbances, and crude oil feed changes in the refinery processes may increase the possibility of outliers in the data set. It should also be pointed out that all of the samples, deemed to be outliers by a certain model, are not necessarily outliers. When multiple outliers are found in the data set, they may mask other outliers, and/or cause clean data point to be regarded as outlier values. Hence, to be able to detect and reduce the effect of outlier observations, robust methods are generally preferred in the literature.

In the current study, in order to identify relations between the process variables and to determine a convenient and efficient method for monitoring CDU process, skipped-PCA and MCD+PCA models were employed. Skipped-PCA method consists of two successively employed PCA models, the second of which is applied on the data points free from the outliers detected by the first PCA model. MCD+PCA method, on the other hand, consists of the application of the robust MCD method on the data, and employing PCA on the remaining data points, from which outliers deemed by the MCD method are removed. In the training set, while observation numbers 118, 130, 132, 136, 137 were detected as outliers by both methods, there is a large number of additional outlier observations, which were detected only by the MCD+PCA method. In summary, while 43 outliers (~21% of training set) were detected by MCD method, only 5 outlier observations were detected by PCA method. It is also seen that subtle perturbations in operation conditions can be detected by the MCD+PCA method, but not by the skipped-PCA method. Using PCA for process diagnostics, the column exit temperature of HD reflux (x_5) was found to be the most important single process variable in determining whether the process is in the normal

operating range or not. The common disturbance pattern in the historical data is seen as the perturbation of the correlation of x_5 with other process variables, and both x_5 and x_6, column pressure, taking outlying values. Monitoring the test data shows that the last 10 observations operated out of normal operational conditions. Examining these observations showed that the desalter pressure (x_2) was significantly perturbed.

HD T95 is one of the most important physical properties affecting the refinery profit; hence online prediction of HD T95, which is measured by ASTM D86 method in TUPRAS İzmit Refinery laboratory, is a great assistance to plant operators and engineers in CDU. The dataset including laboratory measurements of HD T95 may include contaminations, which may result from inhomogeneity in experimental equipment, personnel and conditions, and may be revealed as biased measurements, or heteroscedastic variance. Furthermore, contaminated data may not have normal error distribution due to existence of outliers. Robust statistics aim to give reliable results when error terms do not have normal distribution and/or when there are outliers in dataset. If errors come from nonnormal distributions, the results obtained via LS estimators will not be reliable. Robust regression analysis firstly aims to adapt the majority of the data, and then discovers data points having large residuals from the robust solution, and more accurate models can be built giving less weight those data points. For this purpose, various prediction methods, including LS and robust regression methods, were applied to historical CDU process dataset. Moreover, in order to evaluate the predictive performances of LS and various robust regression methods, MC simulations were used.

In MC simulation study, 200 and 10^5 data points were randomly selected from N(0, 1) distributions with a correlation coefficient of 0.3 for training and testing sets, respectively. Also, random error terms are assumed to be N(0, 1) and N(2, 1) distributed for clean and contaminated data, which comprise 10% of the whole set. For the testing, two sets were produced. While one set contains only clean observations, the other one contains 10% contaminations, as in training set. The aim of using clean test set is testing the prediction accuracy, whereas the aim of using contaminated test set is assessing the metrics (RMSE, MAE) to evaluate the quality of predictions under contamination. Results were obtained by repeating MC simulations for 2000 times. When the LS and various robust regression methods were applied to clean dataset, it was observed that

LTS10%+LS, LTS20%+LS, and all the reweighted LTS estimators yield smaller prediction errors, compared to those from the LS model, sole LTS and modified-two-step LTS estimators. Also, it was detected that when trimming percentage in LTS models reach 40-50%, prediction quality gets worse, so it is not recommended to use LTS models with more than 30% trimming. When the LS and various robust regression methods were applied to contaminated dataset, with the purpose of comparing the efficiencies reliability of prediction measures, MAE and RMSE values were computed for different percentiles of ranked prediction errors (100%, 90%, 80%, 70%, 60% and 50%). Results show that RMSE of 70%-90% and the highest absolute prediction errors are better model assessing methods, compared to 100% RMSE, which is the default metric for assessing predictive capability of models in most of the literature, and, LTS10%+LS and LTS20%+LS methods have smaller RMSE values, consistent with the results obtained for clean test data. The similar picture is obtained for MAEs. One may say that MAE can be used as an alternative to using RMSE on all error terms to measure the predictive performance of a model on test data.

In the model prediction section, in order to construct predictive models for the HD T95 value, first LS method was applied to the training set of the process data. The resulting LS model consists of nine process variables, and 15 predictors, including three cross product and three quadratic terms. R^2 statistics of the LS model is equal to 0.49, with MAE and RMSE statistics equal to 3.67 and 4.62 ^{0}C, respectively. When the LS model is applied to the test data, MAE and RMSE statistics were found to be equal to 3.93 and 5.09 ^{0}C. Four observations (2% of training dataset) were found to be regression outliers using the confidence limits on the LS model. Then, robust regression methods, LTS, M-regression and S-regression, on the same training set were applied. In the LTS method, trimming value was selected as 10%, 20%, 30% and %40. Reweighted LTS were applied via detecting outliers at CLs of 95%, 97.5% and 99%, eliminating the outliers and employing the LS method on the remaining data. GM-estimator was applied for three different tuning values of Huber weight function: 1.1, 1.345, 1.5. S-estimator was applied at three different breakdown points, 20%, 30%, 40%, and at a 97.5% CL. To assess the predictive quality of the models, MAE and RMSE values at different percentages of ranked absolute residuals were computed for all model types and compared to find best predictive model. LTS30% with 97.5% CL was selected as the best model for predicting HD T95, and the analyses

were performed on this estimator. For the training set, LTS30% model yielded MAE, RMSE and R^2 values as 3.62 ^{0}C, 4.79 ^{0}C and 0.69. There are 30 outliers detected by LTS30% model, comprising ~15% of training dataset and these outliers were excluded from dataset. Most of the outliers detected by LTS30% model seem to be sampled from the highest and lowest HD measurements. Moreover, more than ten of the outliers are from successive samples of two observations. This implies the possibility of a dynamic effect of contamination in conditions of two successive days. When the LTS30% model was constructed using test data to evaluate predictive performance, MAE and RMSE statistics were found to be equal to 3.75 ^{0}C and 4.84 ^{0}C, and 74% of the absolute values of the prediction residuals were found to be smaller than 5 ^{0}C. The repeatability of the HD T95 changes between 3.15 ^{0}C and 3.9 ^{0}C, while its reproducibility is between 8.48 ^{0}C and 10.9 ^{0}C, according to ASTM D86 method. LTS30% method yields an estimate of 13.0 ^{0}C for the reproducibility, which is slightly higher than the tabulated value, but still acceptable.

When compared with LS method, none of the parameter estimates of LTS30% model has changed signs, and most of the terms have increased their significance. Also, comparing SE values of LS and LTS30% models; it was observed that robust regression has significantly changed the model parameter estimates, such as x_9, x_{22}, $x_2 \times x_3$, $x_{16} \times x_{22}$, and x_3^2. LTS30% model identified eight leverage points, which LS model could not identify, and these data points might possibly perturb parameter estimates. Furthermore, almost all the absolute residuals above ~7.5 ^{o}C were found to be higher in the LTS30% regression. In order to see the difference in prediction errors clearly, medians of prediction errors of test data were calculated as 1.46 for LS and 0.04 for LTS30% model. From this perspective, it can be concluded that prediction errors from LTS30% model is not biased, while LS model predictions are ~1.5 ^{o}C biased, possibly due to contamination bias in the training set. These all results show that LTS30% model is a more reliable model for HD T95 prediction, and can be used for necessary operational interventions for efficient operation.

Recommendations regarding the present work can be classified in two distinct groups: monitoring and model prediction. For monitoring, other robust multivariate methods, such as Minimum Volume Ellipsoid (MVE), S and MM estimators can be used to detect the sampling periods, in which normal operating conditions are perturbed, and

outliers. To reduce the dimensions of the process, PLS, which is a technique generalizing and combining features from PCA and multiple regression, can be used. Beside PCA and PLS, there are various different methods, such as ANN, Neuro-Fuzzy Systems and Support Vector Machines, which can also be applied to CDU process data. To increase the prediction accuracy, various other robust regression methods, such as R-Estimators, MM-Estimators, Theil-Sen Estimators, can be used. Furthermore, other process variables and lagged observations of HD T95 values (as in an autoregressive model) may be included in the predictive models. In refinery processes, there is large number of quality variables. Hence, robust predictive models may also be constructed for other quality variables, such as flash point, T5 and T95 of products, viscosity, and penetration. Another future study is to employ Robust Principal Component Regression (PCR) and Robust PLS methods. These methods reduce the dimension of multivariable regressors and regress with predicted variable. By this way, there is no need to select regressors as done in LS method with stepwise regression.

REFERENCES

1. Roussel, J., and Boulet, R. (1995b) ‘‘Composition of Crude Oil and Petroleum Products,’’ Chapter 1, In ‘‘Crude Oil Petroleum Products Process Flowsheets,’’ Petroleum Refining, Vol. 1, Wauquier, J. ed., TECHNIP, France

2. Gary, J., Handwerk, G. and Kaiser, M. , *Petroleum Refining:Technology and Economics*, 5th edition, CRC Press, Boca Raton,2007.

3. Worrell, E. and Galitsky, C., *Profile of the petroleum refining industry in California*. [Berkeley, Calif.]: Energy Analysis Dept., Environmental Energy Technologies Dept., Ernest Orlando Berkeley National Laboratory, 2004.

4. Waheed, M. and Oni, A., “Performance improvement of a crude oil distillation unit”, *Applied Thermal Engineering*, 75, pp.315-324, 2015.

5. Liu, J., R., Srinivasan and P.,N., Selvaguru, “Practical problems in developing data-driven soft sensors for quality prediction”, *18th European Symposium on Computer Aided Process Engineering*, Lyon, 2008, Elsevier Science

6. Oracle, *Improving Oil and Gas Performance with Big Data,* 2015, http://www.oracle.com/us/technologies/big-data/big-data-oil-gas-2515144.pdf, [Accessed at October 2016].

7. Martin, E., Morris, A. and Zhang, J., “*Process performance monitoring using multivariate statistical process control*”, IEE Proceedings - Control Theory and Applications, 143(2), pp.132-144, 1996.

8. Sliskovic, D., R., Grbic and Z. Hocenski, “Method for plant data-based process modeling in soft-sensor development”, *Automatika*, 52(4), pp.306-318, 2011.

9. Smith, R., Interpretation of inorganic data, Genium Pub. Corp., Amsterdam, 2001

10. de Smith, M. J., Statistical Analysis Handbook, The Winchelsea Press, Winchelsea, 2015.

11. Kadlec, P., Gabrys, B. and Strandt, S., "*Data-driven soft Sensors in the process industry*", Computers and Chemical Engineering, 33(4), pp.795-814, 2009.

12. Itl.nist.gov, *Linear Least Squares Regression*, 2010, http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd141.htm, [Accessed at October 2016].

13. Cerebro.xu.edu, *Method of Least Squares*, 2016, http://cerebro.xu.edu/math/Sources/Least_Squares/index.html, [Accessed at October 2016].

14. Abdi, H., *The method of least squares*. In N.J. Salkind, D.M., Dougherty, and B. Frey (Eds.): Encyclopedia of Research Design. Thousand Oaks (CA): Sage. pp. 705-708, 2010.

15. Brown, S., "Multiple Linear Regression Analysis: A Matrix Approach with *MATLAB", Alabama Journal of Mathematics, 2009.*

16. *Van de Geer, S., Least square* estimation In Everitt, B. And D., Howell (Eds): Encyclopedia of Statistics in Behavioral Science, Vol.2, , Chichester: Wiley, pp. 1041-1045, 2005.

17. Kano, M. and M. Ogawa, "The state of the art in chemical process control in Japan: Good practice and questionnaire survey", *Journal of Process Control*, 20(9), pp. 969-982,2010.

18. Hubert, M., P., J., Rousseeuw and S., V., Aelst, "High-breakdown robust multivariate methods", *Statistical science*, 23(1), pp. 92-119, 2008

19. Westad, F., *Monitoring chemical processes for early fault detection using multivariate data analysis methods*, 2016, http://www.camo.com/downloads/Monitoring_%20chemical_processes.pdf,[Accessed at November 2016].

20. Brereton, R., "Principle component analysis: Basic ideas", *Alchemist,Chemweb*, 2000.

21. Lu, B., I., Castillo, L., Chiang and T., Edgar, " Industrial PLS model variable selection using moving window variable importance in projection", *Chemometrics and Intelligent Laboratory Systems*, 135, pp.90-109, 2014

22. Sliskovic, D., R., Grbic and Z., Hocenski, "Multivariate statistical process monitoring", Technical Gazette, 19 (1), p. 33, 2012

23. Muthukrishnan, R., Boobalan, E. and Mathaiyan, M., "MCD based principal component analysis in computer vision", *International Journal of Computer Science and Information Technologies*, 5(6), pp. 8293-8296, 2014.

24. Bhar, L. *Robust regression*, 2016, http://www.iasri.res.in/design/ebook/EBADAT/3Diagnostics%20and%20Remedial%20Measures/5-ROBUST%20REGRESSION1.pdf, [Accessed at November 2016].

25. Filzmoser, P. and V., Todorov, "Review of robust multivariate statistical methods in high dimension", *Analytica Chimica Acta*, 705(1-2), pp.2-14, 2011.

26. Wisnowski, J., W., *Multiple outliers in linear regression: Advances in detection methods, robust estimation and variable selection*, Phd, Arizona State University, 1999.

27. Rousseeuw, P., "*Least Median of Squares Regression", Journal of the American Statistical Association*, 79(388), pp. 871-880, 1984.

28. Jacoby**,** B. *Regression III: Advanced Methods*.

29. Andrews, D. and Hampel, F., *Robust estimates of location: survey and advances*. Princeton, N.J.: Princeton, University Press, 1972.

30. Hampel, F., R., "Beyond location parameters: Robust concept and methods", *Bull. Intern. Statist. Inst.,* 46, pp. 375-382, 1975.

31. Rousseeuw, P. J. and A., M., Leroy, *Robust regression and outlier detection*, Wiley-Interscience, New York, 1987.

32. Andersen, R., Modern methods for robust regression, SAGE publications, pp.47-70, 2008.

33. Rousseeuw, P. and A., Leroy, *Robust regression and outlier detection*, New York: Wiley, 1987

34. Doornik, J., "Robust Estimation Using Least Trimmed Squares", 2011.

35. Wang, F. and C. Lee, "An M-Estimator for Estimating the Extended Burr Type III Parameters with Outliers", *Communications in Statistics - Theory and Methods*, 40(2), pp.304-322. 2010.

36. Mathworks, *robustfit*, https://www.mathworks.com/help/stats/robustfit.html, [Accessed at November 2016].

37. Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators.Robust and Nonlinear Time Series Analysis (J. Franke, W. Ha¨rdle and R. Martin, eds.). Lecture Notes in Statist. 26 256–272.Springer, New York

38. Toka, O. and M., Cetin, "The Comparing of S-estimator and M-estimators in Linear Regression", *Gazi University Journal of Science*, 24(4), pp. 747-752, 2010.

39. Rousseeuw, P. and K., Van Driessen, “A Fast Algorithm for the Minimum Covariance Determinant Estimator”, Technometrics, 41(3), p.212, 1998.

40. Colwell, R., F., *Oil Refinery Processes: A brief overview*, Process Engineering Associates, LLC, 2009.

41. MathPro, *An Introduction to Petroleum Refining and the Production of Ultra low Sulfur Gasoline and Diesel Fuel*, 2011, http://www.theicct.org/sites/default/files/publications/ICCT05_Refining_Tutorial_FINAL_R1.pdf, [Accessed at November 2016].

42. Doherty, M., Fidkowski, Z., Malone, M. and Taylor, R., *Perry's chemical engineers' handbook*. 8th ed. New York: McGraw-Hill, 2008.

43. Halvorsen, I., J., *Minimum Energy Requirements in Complex Distillation Arrangements,* Phd, Norwegian University of Science and Technology, 2001.

44. Jukic, A., *Petroleum refining: distillation*, 2013, https://www.fkit.unizg.hr/_download/repository/PRPP_2013_Refinig_dis.pdf, [Accessed at November 2016].

45. ASTM Standard D86-95, 1995, “Standard test methods for distillation of petroleum products”, ASTM International, West Conshohocken, PA, 1995.

46. ASTM Standard D86-15, 2015, “Standard Test Method for Distillation of Petroleum Products and Liquid Fuels at Atmospheric Pressure”, ASTM International, West Conshohocken, PA, 2015.

47. EU Science Hub. *FSDA Matlab code - EU Science Hub - European Commission*, 2015 https://ec.europa.eu/jrc/en/scientific-tool/fsda-matlab-code, [Accessed at December 2016].

48. Diana, G. And C., Tommasi, “Cross-validation methods in principal component analysis: a comparison”, Statistical methods and application, 11, pp. 71-82, 2002.