Use of Multivariate Statistical Techniques in HACCP Programs

by

Umut Başak BALIKLI

**A Dissertation Submitted to the
Graduate School in Partial Fulfillment of the
Requirements for the Degree of**

MASTER OF SCIENCE

**Programme : Food Engineering
Major : Food Engineering**

**İzmir Institute of Technology
İzmir, Turkey**

September, 2003

We approve the thesis of **Umut Başak BALIKLI**

Date of Signature

-- **05.09.2003**

Asst. Prof. Figen (KÖSEBALABAN) TOKATLI
Supervisor
Department of Food Engineering

-- **05.09.2003**

Asst. Prof. Handan ERTÜRK
Department of Food Engineering

-- **05.09.2003**

Asst. Prof. Durmuş ÖZDEMİR
Department of Chemistry

-- **05.09.2003**

Asst. Prof. H. Murat GÜNAYDIN
Department of Architecture

-- **05.09.2003**

Prof. Şebnem HARSA
Department of Food Engineering

-- **05.09.2003**

Prof. Şebnem HARSA
Head of Department

ACKNOWLEDGEMENTS

İ would like to express my sincere gratitude to my advisor, Asst. Prof. Figen (KÖSEBALABAN) TOKATLI for her supervision, guidance and encouragement throughout this study.

I am also grateful to Prof. Velittin GÜRGÜN and Melekber COŞKUN for their help and technical support.

Special thanks to Emre YALAMAÇ for his support, encouragement and understanding.

I would like to appreciate deeply to Dilek DEMİRBÜKER, Berna ÜZELYALÇIN and all my friends for their friendship, support and understanding.

Finally, I would like to thank my family for their support and encouragement.

ABSTRACT

Food safety is the major concern for the food industry and Hazard Analysis and Critical Control Points (HACCP) is an effective safety management system. Data analysis is an important ingredient of this system. The use of Statistical Process Monitoring (SPM) methods in critical control point monitoring step can further improve a HACCP system, since SPM and HACCP have a common goal which is to prevent failures before they occur. Food production processes include many variables and generally they are not independent of each other. The use of multivariate statistical methods is more appropriate than that of univariate statistical methods for food processes and provides comprehensive analysis of the data. The aim of this study was to display the benefits of the use of multivariate SPM techniques in HACCP system.

In this study, data were taken from a food processing plant, which uses HACCP program in the production. They were collected in a frozen vegetable production line and composed of raw material properties, process conditions, microbiological counts and end product analyses. The data were analyzed by using multivariate statistical techniques such as Principal Component Analysis (PCA), Multiple Linear Regression (MLR), Principle Component Regression (PCR) and Partial Least Square Regression (PLSR). In the monitoring step, multivariate statistical tools such as Hotelling's T^2, Squared Prediction Error (SPE) and contribution plots were utilized. Cause and effect diagrams were also employed as a problem analysis tool to improve the process.

Uncorrelated score variables of PCA of process data and quality data successfully analyzed out of control observations on time basis in T^2 and SPE plots. Contribution plots displayed the responsible variables, which alarmed at particular time instant. Contribution percentages of variables obtained from these out of control points displayed that blanching temperature and microbial counts are very important contributing factors. Blanching temperature is a variable of the first critical control point (CCP-1) and microbial counts are the verification of that CCP. This result indicates that CCP-1 is the point which extra care should be taken.

PCR and PLSR techniques were successful in analyzing the process and product data individually. T^2 and SPE plots of these models were nearly the same with the PCA of

process data and product data. The regression models (MLR, PCR and PLSR) were not able to explain the correlation structure between process and product data, completely. The in-control data set used in this study was insufficient to construct regression models since it failed to explain the normal operating conditions exactly.

It was stated that the proper data collection in the production line would cause an enhancement in the application of multivariate statistical techniques, in both monitoring and prediction of critical control point measurements.

ÖZ

Gıda güvenliği, gıda endüstrisi için en önemli konudur ve Tehlike Analizi ve Kritik Kontrol Noktaları (HACCP) etkili bir güvenlik yönetimi sistemidir. Veri analizi bu sistemin önemli bir parçasıdır. İstatistiksel Süreç Gözleme (İSG) ve HACCP sisteminin her ikisinde de ortak amaç hataların oluşmadan önlenmesi olduğundan, kritik kontrol noktalarının izlenmesi aşamasında İSG yöntemlerinin kullanımı HACCP sistemini daha etkili kılar. Gıda üretim süreçleri çok sayıda değişken içerir ve bu değişkenler genellikle birbirinden bağımsız değillerdir. Bu nedenle gıda üretiminde çok değişkenli yöntemlerin kullanımı tek değişkenli yöntemlere göre daha uygundur ve kapsamlı bir veri analizi sağlar. Bu çalışmanın amacı çok değişkenli istatistiksel süreç izleme tekniklerinin HACCP sisteminde kullanılmasının faydalarını göstermektir.

Bu çalışmada kullanılan veriler üretimlerinde HACCP programı uygulayan bir gıda fabrikasından alınmıştır. Dondurulmuş gıda hattında toplanan hammadde özellikleri, üretim koşulları, mikrobiyolojik sayımlar ve son ürün analizleri kullanılan veri setlerini oluşturmuştur. Toplanan veriler, temel bileşenler analizi (PCA), çoklu doğrusal bağlanım denklemi (MLR), temel bileşenler bağlanım denklemi (PCR) ve kısmi en küçük kareler yöntemi (PLSR) gibi çok değişkenli istatistiksel teknikler ile analiz edilmiştir. İstatistiksel süreç izleme araçları olarak Hotelling T^2 çizimi, öngörme hatalarının karesi (SPE) ve katılım grafikleri kullanılmıştır. Süreci geliştirmek amacıyla problem analizi aracı olarak neden ve etki çizelgelerine de yer verilmiştir.

Üretim, hammadde ve son ürün verilerinin PCA analizine ait bağıntısız skor değişkenleri, kontrol dışı ölçümleri T^2 ve SPE çizimlerinde zaman bazında başarıyla analiz etmiştir. Katılım grafikleri, belirli bir zamanda alarm veren sorumlu değişkenleri belirlemiştir. Değişkenlerin kontrol dışı noktalardan elde edilen katılım yüzdeleri, haşlama sıcaklığı ve mikrobiyal sayımların önemli katılım faktörleri olduğunu göstermiştir. Haşlama sıcaklığı ilk kritik kontrol noktasına (CCP-1) ait bir değişkendir ve mikrobiyal sayımlar bu kritik kontrol noktasının doğrulamasıdır. Bu sonuç CCP-1'in özellikle dikkat gösterilmesi gereken bir nokta olduğunu göstermektedir.

PCR ve PLSR teknikleri, üretim, hammadde ve son ürün verilerinin ayrı ayrı analizinde başarılı olmuştur. Bu modellerin T^2 ve SPE çizimleri, üretim, hammadde ve son

ürün verilerinin PCA analizi sonuçlarının yaklaşık aynısıdır. Bağlanım denklemi modelleri (MLR, PCR ve PLSR), üretim, hammadde verileri ve son ürün verileri arasındaki bağıntılı yapıyı tam olarak açıklayamamaktadır. Bu çalışmada kullanılan kontrol-içi veri seti normal operasyon koşullarını açıklamada başarısız olduğundan, bağlanım denklemi modellerinin oluşturulması için yetersizdir.

Üretim hattında uygun veri toplanmasının, kritik kontrol noktalarındaki ölçümlerin gözlenmesi ve öngörülmesinde çok değişkenli istatistiksel tekniklerin uygulanmasını zenginleştireceği belirlenmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

a : latent variable index

c : contribution matrix

C : total contribution matrix

e : error vector

E : residual (error) matrix of **X**

F : residual (error) matrix of **Y**

$\mathbf{F}^*$: residual (error) matrix of **Y**

$F_{\alpha,p,n-p}$: F distribution with p and n-p degrees of freedom

j : variable index

k : observation index

n : number of observations

p : number of variables of **X**

q : number of variables of **Y**

p : loading vector

P : loading matrix of **X**

Q : loading matrix of **Y**

r : number of latent variables

$\mathbf{S}_e$: estimate of covariance matrix of in-control errors

$\mathbf{S}_\mathbf{X}$: estimate of covariance matrix of in-control data

SPE : squared prediction errors

t : score vector

T : score matrix of **X**

U : score matrix of **Y**

$\mathbf{T}^2$: Hotelling's T^2 values

W : weight matrix

X : process variables (independent variables) data matrix

$\hat{\mathbf{X}}$: predicted data matrix

Y : quality variables (dependent variables) data matrix

α : significance level

β : regression coefficients matrix

λ : eigenvector

Chapter 1
INTRODUCTION

Food safety is the primary concern of the food industry because of its importance for public health, trade and economy (Barendsz, 1998). Foodborne outbreaks are still being observed confirming the importance of safety concerns in spite of the progress in food science and technology (Jouve *et al*., 1999). Therefore, the quality concept in the food industry should be based on safety. The demand for safer foods and regulations of modern trading conditions require the utilization of safety or quality management systems in food production. Hazard Analysis and Critical Control Points (HACCP) is a food safety management system, which is directed to ensure the safety of the produced food. HACCP identifies and monitors specific foodborne hazards that can adversely affect the safety of the food product. The system has some basic principles and monitoring of Critical Control Points (CCPs) is one of them. The purpose of the monitoring principle is to observe the important parameters in order to evaluate whether the CCP is in control.

Statistical Process Monitoring (SPM) is a powerful tool for achieving the objective which is to keep manufacturing process stable, to improve process performance and to reduce variability on key parameters by using control charts (Montgomery, 2000). Since SPM and HACCP are both prevention-based procedures, they employ regular evaluation of the process steps instead of the final inspection to make sure that the complete process is in control. If a process step is not in control, then corrective action is taken immediately. Thus, product quality is assured through a controlled process (Does *et al*., 1999).

Even the use of SPM in HACCP system is not obligatory, it is extremely important that the monitoring procedures applied to CCP's are to be statistically valid and monitoring is the most beneficial when established under a system of Statistical process monitoring in practice (Mortimore and Wallace, 1998). HACCP and SPM integration could further enhance the prevention philosophy by data analysis, which displays warnings for out of control cases, and defects can be prevented with appropriate corrective actions when the process is out of control. SPM integration improves the HACCP system since eliminating potential risks increases the safety of the product. The use of SPM in HACCP system also leads to a process within critical limits and provides verification. Therefore, SPM provides a more effective HACCP system.

SPM employs many techniques and methods in order to control and improve the process. These techniques can be univariate and multivariate. Large amount of correlated data is collected in modern industrial operations and statistical analysis is inevitable. Univariate techniques are not capable of handling large complex data sets whereas multivariate SPM methods analyze large sets of correlated data. Likewise many others, food production is a multivariable process and includes various correlated variables. Thus, the use of multivariate SPM for the food production process within the HACCP system would be reasonable.

The objective of this study is to display the advantages of the use of multivariate SPM procedures in HACCP system. These multivariate methods evaluate the process variables simultaneously. The data collected from a food processing plant applying HACCP program were analyzed by using multivariate SPM techniques to determine the out of control cases due to special causes. Once these out of control cases in the process are determined, the occurrence of corresponding special causes can be prevented and this prevention also reduces the need to test the finished product. Therefore, the efficiency of the HACCP system can be increased. The most significant point of this study is that the data used were from an industrial process, not obtained by an experimental study or by a computer simulation.

In this study, the process is modeled by using the multivariate statistical methods, which are Principal Component Analysis (PCA), Multiple Linear Regression (MLR), Principal Component Regression (PCR) and Partial Least Square Regression (PLSR). It is monitored by using the multivariate statistical monitoring tools, which are Hotelling's T^2 plot and Squared Prediction Error (SPE) chart. Contribution Plots and Cause and Effect Diagrams are also employed to diagnose the causes of faults in the process.

Chapter 2
FOOD SAFETY AND HACCP

2.1. Food Safety

Food safety is a major problem for the food industry not only because of its importance for public health but also because of its impact on trade and its economic significance (Barendsz, 1998). Foodborne outbreaks, which are still being observed despite the progress in medicine, food science and technology of food production, prove that the safety concern is not irrelevant. According to the reports from 11 European countries, 120 cases of foodborne illness per 100,000 people occurred during the year 1990. Also in some European countries, nearly 30,000 cases of acute gastroenteritis per 100,000 people each year are estimated to be foodborne. Obviously, not all of the foodborne illness cases reached food inspection, control and health agencies (Jouve *et al*., 1999).

Food safety is gaining importance on a global scale because of recent well-known foodborne disease outbreaks. Figure 2.1 shows the contributing factors of confirmed foodborne outbreaks in US between 1988-1992. The factors shown in the figure are the points, which can be eliminated if controlled properly. Thus, all companies and organizations involved in food production should consider safety concerns first. Safety concerns also have a significant role in national and international legislation and trade. Since food production has a direct health risk to the consumer, many countries have their legislation and regulations for food processing requiring food processors to ensure the safety of their products. Legislation also affects national and international trade and competition between companies. Therefore, economy of companies and in large-scale, economy of countries are directly effected by food safety concerns.

All these factors challenge the food industry to ensure safety through an appropriate management programme. Such a programme should guarantee that the food company increases its commitment to product safety to the highest level and should focus on improvements that can be applied to both organizational and technical issues (Jouve *et al*., 1999). Therefore, the quality concept in the food industry should primarily include safety considerations.

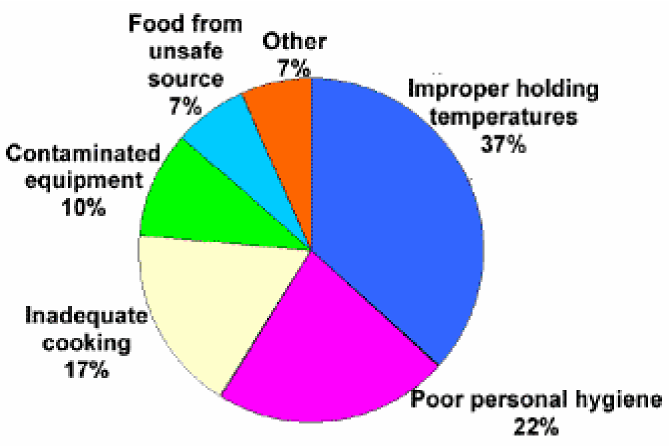


Figure 2.1. Contributing Factors of Confirmed Foodborne Outbreaks in US 1988-1992 (US Department of Health & Human Services, Public Health Service, Morbidity and Mortality Report, Surveillance for Foodborne Disease Outbreaks--US, 1988-1992, Vol. 45, No. SS-5, October 25, 1996)

2.2. HACCP

The demand for safer foods and regulations of modern trading conditions necessitate the widespread use of safety or quality management systems such as Hazard Analysis and Critical Control Points (HACCP), ISO-9000 Series of Standards and Total Quality Management (TQM) in the food industry. ISO-9000 and TQM are the management systems, which assure quality. TQM defines a long-term managerial strategy and involves other management systems such as ISO-9000. HACCP is the assurance of safety and specific to process. It differs from quality management systems with this property. Thus, it generally takes part under ISO-9000 and TQM (Jouve *et al.*, 1999).

HACCP is a food safety management system, which aims to ensure the safety of the produced food (Topal, 2001). Food Safety and Inspection Service (FSIS) defines HACCP as "a way for industry to control and prevent problems and ensure safe food by controlling the production process from beginning to end, rather than detecting problems at the end of the line". HACCP is a prevention-based food safety system, which identifies and monitors specific foodborne hazards that can adversely affect the safety of the food product. The importance of HACCP comes from its effectiveness on maximizing product safety (Mortimore and Wallace, 1998). HACCP first appeared at the beginning of the 1960’s and was used to produce 100% safe foods for NASA astronauts to consume during space flights. HACCP system was first widely applied to low acidic canned foods in industry during the 1970's and the use of HACCP system to ensure food safety has increasingly spread in the food industry (Arıkbay, 2002). A hazard is a biological, chemical or physical property that will cause a food to be unsafe for consumption and a Critical Control Point (CCP) is a point in a process at which control can and should be applied to prevent, eliminate or minimize a potential food safety hazard. HACCP provides a systematic method to analyze the food process and identifies potential biological (e.g. cross-contamination, pathogens), chemical (e.g. allergens, cleaners, residues, natural toxins) and physical (e.g. glass, metal, foreign objects) hazards that can occur in food. HACCP systems are designed to prevent the occurrence of these potential food safety problems. This is achieved by assessing the inherent risk attribute to a product or a process and by determining the necessary steps that will control the identified risks. In addition, HACCP requires the development of strategies to reduce these hazards to an acceptable level in food (Surak *et al*., 1998). In this manner, potential problems are determined at an early stage in a food production. Preventing problems from occurring is the superiority of HACCP system when compared to other approaches such as inspection and end-product testing (Ehiri *et al*., 1995).

HACCP, which is the most effective means of managing food safety, is increasing on a worldwide basis and implementation of HACCP accelerates as it becomes a regulatory requirement. The difficulty in focusing on legislation about food safety is that legislation is ever-changing and extending its scope. In the United States, the implementation of HACCP in low acid canned foods, all meat and poultry production and seafood production is required by law since 1990's. The trend seems that HACCP will be mandatory for all US food processing facilities. In Canada, HACCP is required for high-risk food products processing since 1991. European Community is also very strict on food safety issues. European Community Directive 93/43 EC (1993) does not use the precise wording of Codex Alimentarius or National Advisory Committee on Microbiological Criteria for Food (NACMCF) for HACCP but includes HACCP principles in basis. The Directive strongly recommends all food businesses throughout Europe to use HACCP approach. In summary, it is clear that legislation of all developed countries tends to make HACCP a mandatory requirement for the food industry (Mortimore and Wallace, 1998). Similar to the European Community, Turkey also does not use the exact wording of Codex Alimentarius or NACMCF for HACCP but includes HACCP principles in basis in food law (Resmi Gazete, 2002). Turkey's food law introduces the hygienic production and control of food products with the principles of HACCP and emphasizes their necessities. However, it does not force all food industries to use HACCP principles. Priorities on this issue are the processors of meat, milk and water. They should apply the principles within certain time periods changing from 2 to 10 years. Frozen food processors are not responsible from the application of these principles yet.

HACCP system is unique to process and its function is to utilize scientific methods to monitor process performance and improvement. National Advisory Committee on Microbiological Criteria for Food (NACMCF) defined the preliminary tasks in the development of the HACCP plan and HACCP principles as in Table 2.1.

Table 2.1. Preliminary Tasks in the Development of the HACCP Plan and HACCP Principles (NACMCF–National Advisory Committee on Microbiological Criteria for Food, 1997)

Task 1: Assemble the HACCP Team
Task 2: Describe the food and its distribution
Task 3: Describe the intended use and consumers of the food
Task 4: Develop a flow diagram that describes the process
Task 5: Verify the flow diagram
Principle 1: Conduct a hazard analysis
Principle 2: Determine the critical control points (CCPs)
Principle 3: Establish critical limits
Principle 4: Establish monitoring procedures
Principle 5: Establish corrective actions
Principle 6: Establish verification procedures
Principle 7: Establish record-keeping and documentation procedures

It is important to establish and implement the HACCP system with a multi-disciplinary team effort and the HACCP team should include qualified personnel from different departments of the company such as quality assurance, operations/production, engineering etc. The team should describe the food, its intended use, consumer and distribution. It is essential to obtain the best understanding of the process for the

HACCP team. Therefore, the team should develop a flow diagram that describes the process and verify the diagram in place.

The first principle of HACCP, which the HACCP team should cover, is to conduct a hazard analysis. This is done by preparing a very detailed list of process steps and analyzing them for possible physical, chemical and biological hazards. Description of the control measures for these hazards also takes place in this section. Second principle is to determine the Critical Control Points where control is critical for assuring the product safety. Each CCP is displayed with its number and the capital of its potential hazard (e.g. CCP-3B means this process step is the third CCP and the potential hazard at this point is biological). Then, critical limits are established for each CCP in measurable units at the third step. The HACCP team should specify monitoring procedures to control whether a particular CCP is within its critical limits or not. Identification of monitoring actions, frequency and responsibility constitute Principle 4. Corrective actions are the measures, which should be employed when monitoring indicates that a particular CCP is not under control and the subject of Principle 5. Principle 6 is to establish verification procedures, which confirm that the HACCP system is working correctly. The last step is establishment of record-keeping and documentation procedures in order to indicate the HACCP system is operating under control. Documentation should include any deviations from critical limits and the corrective actions that have been taken. Performing all of the necessities of HACCP system mentioned above, proves safe food manufacture (Mortimore and Wallace, 1998).

Monitoring, which is the subject of Principle 4, includes a sequence of observations or measurements of control parameters of a CCP to assess whether it is in-control. Unfortunately, even if data are collected to monitor a CCP, this step of HACCP system is limited with simply checking over the data according to their critical limits by the companies in most of the time. Microbial testing is often used as a monitoring tool but it is not feasible since this method can not prevent a failure at an early stage of the process and end product testing can not determine the root cause of a safety problem. Microbial testing should be used as a verification procedure in a HACCP system.

These approaches above lead to fail in to detect process changes over time and to capture a problem before exists since the data collected are not used effectively (Surak *et al.*, 1998). However, it is extremely important that the monitoring procedures applied to CCP's are statistically valid and monitoring is most beneficial when established under a system of statistical process monitoring (Mortimore and Wallace, 1998; Hayes *et al*., 1997; Surak *et al*., 1998). Although the purpose of HACCP programs is food safety issues, they also enhance the quality of products. Likewise, quality control and quality improvement processes provide benefits to HACCP by reducing the risk of a food safety hazard. Thus, statistical process monitoring techniques, quality management systems and all kinds of process improvement and problem solving tools are beneficial for HACCP Programs. All HACCP authorities such as Codex Alimentarius and NACMCF also advice the use of SPM within HACCP system.

Chapter 3
STATISTICAL PROCESS MONITORING

3.1. Statistical Thinking and SPM

It is certain that variation exists in all kind of processes. Understanding and reducing variation are the key factors of statistical thinking. There are two types of variation;

- Variation from common causes
- Variation from special causes

Common causes are the causes of variation, which present all the time in a process and have a small effect on the variation individually. It is not possible to remove common causes. Special causes (assignable causes) do not always present in a process but arise from outside the usual process and have a much bigger impact on variation than any single common cause. For example, in the measurements of a probe surrounding temperature fluctuations lead to a common cause of variation. However, a calibration problem of that device results in a special cause of variation. A process that has only common cause of variation is said to be stable or predictable and if a process is stable or predictable, it is in statistical control. If a process has special causes of variation, it is not in statistical control.

The data from the process should be statistically analyzed in order to understand the causes of variation. The most common way for this aim is to plot the data in time order. Such plots are called statistical control charts and include statistical control limits. If a measurement that belongs to the process is within control limits, it is assumed that the variation is random (i.e. from common causes). In such a case, the process is stable, in statistical control and immediate future is predictable. A point outside the control limits is a signal of a special cause, which indicates the need for action. The reaction in case of special cause variation is to discover the recent differences on the system and take corrective action to prevent reoccurrence (Montgomery, 2000; Surak *et al*., 1998; Does *et al*., 1999).

Managing, controlling and reducing variation are the major purposes of quality management. Since they are also included in statistics, variation is the actual link between quality management and statistics. This relation leads to the wide use of statistical methods in quality problems. Statistical Process Monitoring (SPM) is a procedure that uses control charts and other graphical problem solving tools to provide an effective process control. When used effectively, SPM can be a powerful tool for process improvement (Miller and Balch, 1991). Process control based on SPM provides defect prevention instead of defect detection. Therefore, SPM is process-oriented instead of product-oriented. In a production process controlled with SPM, and also in HACCP, final inspection is replaced with regular evaluation of the several process steps to make sure that the complete process is still functioning normally. If a process step is found to be no longer functioning properly, action is taken immediately to correct it. In this way, product quality is assured through a controlled process (Does *et al.*, 1999).

3.2. SPM Techniques

SPM, which is an important ingredient of control and monitoring of systems, includes a large number of techniques and methods directed to control and improvement of the process. SPM techniques can be divided into two according to the size of the process data or correlation between the process variables. They are univariate and multivariate methods.

3.2.1. Univariate SPM Techniques

Univariate SPM methods monitor univariate problems or sets of independent variables by using control charts. Univariate techniques need no model and use the data directly since they accomplish with uncorrelated data. Control charts are simple representations of a quality characteristic plotted in time. Therefore, they can be easily applied and interpreted. Control limits determined by the mean and the variance of an in-control data set provide monitoring the stability of the process and the presence of special causes (Figure 3.1).

Shewhart (X-bar and Range), Cumulative sum (CUSUM) and Exponentially Weighted Moving Average (EWMA) Charts are the examples of statistical tools which can be used for univariate problems. The type of assignable cause(s) that can potentially affect the process should be considered when selecting the type of control chart, which is to be used for process monitoring (Runger and Montgomery, 1997; Montgomery, 2000).

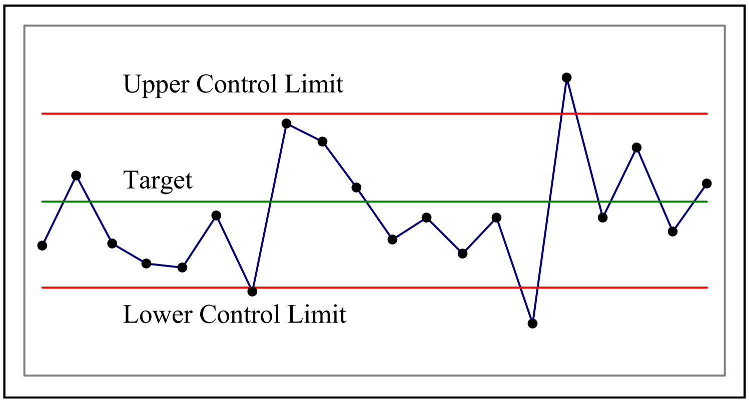


Figure 3.1. Sample Univariate Chart

There are two major disadvantages of univariate methods. The first one is the assumption of independence of observations. In other words, it is assumed that the data are not autocorrelated when univariate charts are used. Autocorrelation is correlation of values that are adjacent in time. Univariate charts give misleading results and too many false alarms if the data are autocorrelated (Montgomery, 2000). The limits of the univariate charts are determined for a univariate data set. However, control limits change when the correlation of variables is taken into account. The second disadvantage is their inability to monitor multiple variables (Kresta *et al.*, 1991). Univariate techniques do not consider the correlation within variables since they are generated separately for each variable.

3.2.2. Multivariate SPM Techniques and Tools

In the conditions of today's modern industry, overwhelming amount of data is collected in many processes and they are generally dependent to each other. In order to obtain the most useful information about the process, the data should be analyzed statistically. Since the univariate techniques are not sufficient for large complex data sets, multivariate techniques should be utilized. Multivariate SPM methods are suitable to analyze large sets of correlated data. They use complete data set by taking into account the correlation between the variables. Multivariate SPM techniques provide efficient simplification and interpretation of many different variables simultaneously (Martens and Russwurm, 1983). The aim is to detect special causes in the process and to improve the process by eliminating these special causes (Miller *et al.*, 1998). The disadvantages of multivariate methods are their difficulty to understand because of their complexity and the difficulty of applying them to the data. However, the complexity and time consuming properties of multivariate methods have disappeared with the advanced computation facility of various software (Mellinger, 1987).

The procedure in multivariate SPM includes several steps. These are;

- to collect or select "in-control" data to form the historical database
- to develop the statistical model that characterize normal operating conditions by using in-control data
- to apply new process data to the model
- to construct control chart with control limits to monitor process operation and product quality
- to declare the process to be out-of-control when the data are outside the control limits

In the following sections, the multivariate statistical modeling and regression techniques PCA, MLR, PCR and PLSR are explained. The multivariate statistical process monitoring charts T^2 and SPE are given, as well.

3.2.2.1. Principal Component Analysis (PCA) :

PCA is a method, which is used for the analysis of a single data matrix. It reduces the variation among many variables into a few latent factors (Figure 3.2). PCA is appropriate for data including numerous variables and where these variables are highly correlated. The idea behind PCA is the use of projection to model high-dimensional data in a low dimensional latent variable subspace that describes most of the variability in the data (Rodriguez and Tobias, 1999).

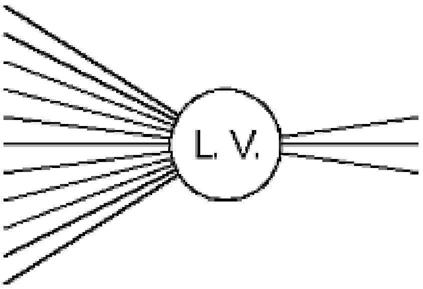


Figure 3.2. Reducing the Original Variables into a Few Latent Variables

These new latent variables summarize all the important information contained in the original data and define the plane of greatest variability (MacGregor *et al*., 1994). **X** matrix has n observations on rows and p variables on columns ($\mathbf{X}_{nxp}$). The model decomposes **X** matrix into r principal components ($PC_1+PC_2+...+PC_r$) and random error ($\mathbf{E}_{nxp}$). PCA algorithm is explained in Appendix A1. The idea behind PCA method is given by the following equations :

$$\begin{aligned} X &= \text{PCA model} + \text{residual} \\ X &= PC_1 + PC_2 + ... + PC_r + E \end{aligned} \tag{3.1}$$

The PCA model in terms of its score and loading matrices is given as :

$$\begin{aligned} \mathbf{X} &= \mathbf{TP'}+\mathbf{E} \\ \mathbf{T} &= \mathbf{XP} \end{aligned} \tag{3.2}$$

X is the nxp data matrix where **T** is nxr matrix of scores, **P** is pxr matrix of loadings. **E** is nxp matrix of errors. n is the number of observations, p is the number of process variables in the data set. r is the number of principal components.

Column vectors of the score matrix (**T**) include the latent variables (or principal components) and these new variables are uncorrelated. The first principal component explains most of the process variation in the data set. The second principal component has the next largest variability and it is perpendicular to the first dimension (Figure 3.3), and so on. PCA provides a reduction in dimension by selecting dimensions causing the most variability and neglecting the remaining sources of variation (Miller *et al*., 1998). Loading matrix (**P**) involves the weights of the original variables in the principal components. Score matrix should be obtained by using the in-control data set and its loading matrix for the model development and this model is applied to the data set. **E** matrix represents the random error, which is the difference between original and predicted values, and its components are independent of each other if the model is properly constructed.

Scree test can be used in order to decide how many PCs to use or the PCs are determined by the eigenvalue rule (eigenvalue >= 1.0). The number of PCs can also be decided according to the percentage of the total variability explained (Johnson and Wichern, 1998; Gonzales *et al*., 2000). Scree plot is a useful tool, which helps to decide the number of principal components retained in the model. Eigenvalues are ordered from largest to smallest and plotted in scree plot (Figure 3.4). λ_i is the magnitude of eigenvalue and i is its number (Johnson and Wichern, 1998).

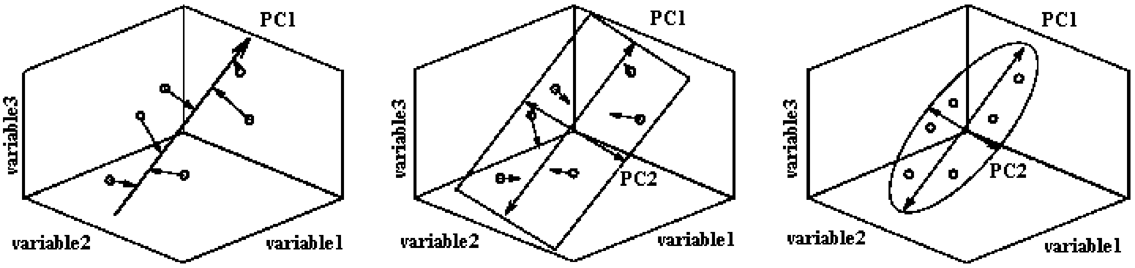


Figure 3.3. Geometric Representation of the Principal Components (Rodriguez and Tobias, 1999)

Plots of the principal component scores and loadings are effective tools for data analysis. Scatter plots of scores are the representation of two score vectors graphed versus each other and shows the projected locations of measurements onto the components. Loading plot is the illustration of two loading components graphed versus each other. It explains how much each variable contributes to a particular PC (Sahni *et al.*, 1999). The major advantage of score plot is that it is possible to determine similar problems which are non-sequential in time sequence since they usually cluster together in this kind of a plot (Miller *et al.*, 1998; Johnson and Wichern, 1998).

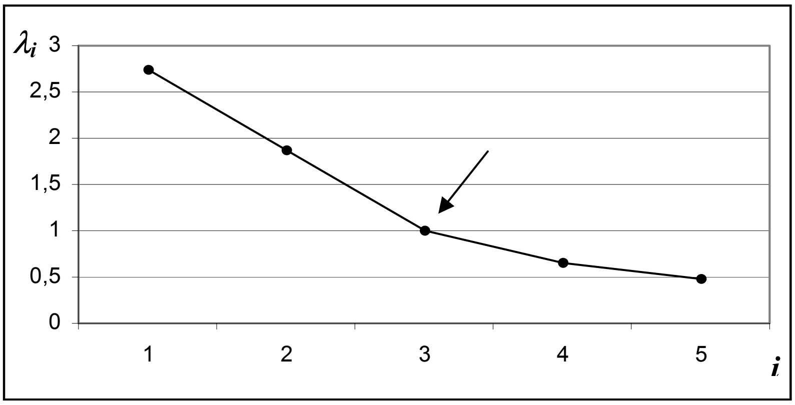


Figure 3.4. Sample Scree Plot

The effect of predictor (independent) variables (**X**) on the response (dependent) variables (**Y**) is assessed by a statistical methodology which is called regression analysis (Johnson and Wichern, 1998). The goal of the regression techniques is mostly the prediction of **Y** by using the information coming from **X** data. Multiple Linear Regression (MLR), Principal Component Regression (PCR) and Partial Least Square Regression (PLSR) are the most common regression techniques used for the analysis of two data matrices; **X** and **Y**.

3.2.2.2. Multiple Linear Regression (MLR) :

MLR is the most common and simplest way to model a linear relationship between response and predictor variables (**Y** and **X** respectively). MLR model :

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{F} \\ \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \end{aligned} \tag{3.3}$$

β which has the dimension of p rows by q columns (β_{pxq}) is the regression coefficient matrix obtained from in-control $\mathbf{X}_{nxp}$ and $\mathbf{Y}_{nxq}$ data matrices by ordinary least squares method. It is used to construct the relationship between these matrices. The model should extract all the information and the remaining part is random error ($\mathbf{F}_{nxq}$). The error term is assumed to be normally distributed with zero mean and constant variance (Johnson and Wichern, 1998).

MLR techniques may have severe problems when applied to correlated data with many variables (Kresta *et al.*, 1991). In the least squares estimation of β matrix (Equation 3.3), the inverse operation of **X'X** matrix can not be performed when the columns of **X** are collinear. Other regression techniques such as PCR and PLSR are employed to reduce the collinearity in **X** matrix.

3.2.2.3. Principal Component Regression (PCR) :

PCR is another regression method. PCR is a two-step multivariate technique: in the first step PCA is applied to the data matrix **X**. The variables of **X** are converted into scores (**T**). This is followed by a MLR step between the scores obtained in the PCA step and the **Y** matrix to be modeled (Maesschalck *et al.*, 1999). PCR overcomes both the dimensionality and collinearity problems since it uses the scores of **X** matrix instead of **X** itself. These scores are orthogonal to each other and reduced in dimension (Kresta *et* *al*., 1991). The algorithm is given by Appendix A2. In terms of $\mathbf{X}_{nxp}$ data matrix and PCA score matrix $\mathbf{T}_{nxr}$, the regression coefficient $\boldsymbol{\beta}$ and **Y** data matrices are :

$$\begin{aligned}
&\mathbf{T} = \mathbf{XP} \\
&\boldsymbol{\beta} = (\mathbf{T'T})^{-1}\mathbf{T'Y} \qquad (3.4) \\
&\mathbf{Y} = \mathbf{T}\boldsymbol{\beta} + \mathbf{F}
\end{aligned}$$

β_{rxq} and $\mathbf{F}_{nxq}$ gives the regression coefficients and error matrices respectively where β is rxq matrix. **F** is nxq error matrix of the PCR model and has the same dimension as **Y**. q is the number of quality or end product variables.

However, PCR has a disadvantage of being two step method. It has the risk that some useful information will end up in discarded principal components and some noise will remain in the components used for regression (Geladi and Kowalski, 1986). Another drawback that may cause lack of predictive ability is that **Y** data are not used in the computation of score variables.

3.2.2.4. Partial Least Square Regression (PLSR) :

PLSR method, which is also known as Projection to Latent Structures, is suitable for handling collinear or highly correlated data and it is a good alternative to the more classical MLR and PCR methods. PLSR appears to best address dimensionality and collinearity problems mentioned above. It is a regression method which extracts latent variables that not only explain the variation in the process data (**X**), but also the variation in **X** which is the most predictive of the corresponding product quality data (**Y**) (Kourti and MacGregor, 1995; Geladi and Kowalski, 1986). There are two main PLSR algorithms named PLS1 and PLS2. In PLSR1, **Y** block has single variable or the model is built for each variable in a multicomponent case. This means that for an n-component sample, n different models have to be built. Components of **Y** are assumed to be independent of each other in PLS1. PLS2 is appropriate for multicomponent **Y** block, as it can model several components simultaneously and the components are assumed to be correlated. PLSR model refers to PLS2 in multivariate SPM since variables of **Y** matrix are correlated in most of the cases. Thus, PLS2 was used in this study. Model development of PLSR is quite similar with PCA. PLSR analyses two data sets, process data ($\mathbf{X}_{nxp}$) and product quality data ($\mathbf{Y}_{nxq}$), in cause and effect point of view where PCA analyses the variability in a single data set. PLSR method includes outer (**X** and **Y** individually) and inner (linking **X** and **Y**) relations of two data matrices. It reduces the dimensions of **X** and **Y** data matrices and find scores for both. Outer relations of **X** and **Y** data are represented in PLSR method as in Equation 3.5.

$$\begin{aligned} X &= \text{PLSR model} + E \\ Y &= \text{PLSR model} + F \end{aligned} \tag{3.5}$$

PLSR model is determined with a common algorithm known as NIPALS (nonlinear iterative partial least squares). The loading matrices and scores of both **X** and **Y** matrices are computed in an iterative manner in which the principal components or factors are determined one at a time. The PLS algorithm is in Appendix A3. As a result of NIPALS algorithm, in terms of loading matrices and scores, **X** and **Y** matrices are :

$$\begin{aligned} \mathbf{X} &= \mathbf{TP'}+\mathbf{E} \\ \mathbf{Y} &= \mathbf{UQ'}+\mathbf{F} \end{aligned} \tag{3.6}$$

$\mathbf{T}_{nxr}$ and $\mathbf{U}_{nxr}$ are the score matrices where $\mathbf{P}_{pxr}$ and $\mathbf{Q}_{qxr}$ are corresponding loading matrices. $\mathbf{E}_{nxp}$ and $\mathbf{F}_{nxq}$ define the error terms (Equation 3.6). PLSR model also consists of a correlation between **X** and **Y** blocks. Inner relation of **X** and **Y** data is represented in PLSR method as :

$$\begin{aligned} \boldsymbol{\beta} &= \mathbf{W}(\mathbf{P'W})^{-1}\mathbf{Q'} \\ \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{F}^{*} \end{aligned} \tag{3.7}$$

β is the regression coefficients matrix of rxq dimension. **W** is a pxr weight matrix. $\mathbf{F}^{*}$ is a nxq residual matrix having the same dimension as **Y**.

There are some common features of PCA, PCR and PLSR;

- to extract new uncorrelated variables (principal components or latent variables) from initial variables
- to reduce the dimension of the data by these new structures (i.e. data reduction)
- to explain the variability with few uncorrelated variables including maximum information of the data

The most important purpose of statistical modeling of the multivariate process data is to monitor the production line by means of statistical charts on time basis. In the following sections, two multivariate monitoring charts and the contributions of variables on the chart statistics are given.

3.2.2.5. Hotelling's T^2 :

The T^2 statistic is a method, which is used to monitor a large number of process variables with a single statistic. This method is used to detect out of control signals in large data sets. T^2 method measures the deviation of a set of variables from their mean values in a certain time instant (Kourti and MacGregor, 1995). The T^2 statistic at time instant k is calculated as :

$$T_k^{\ 2} = \mathbf{x}_k \mathbf{S}_X^{\ -1} \mathbf{x}_k' \tag{3.8}$$

Where T$_k^2$ is a scalar, $\mathbf{x}_k$ is 1xp observation vector of $\mathbf{X}_{nxp}$ data matrix.$\mathbf{S}_X^{-1}$ is the pxp inverse covariance matrix of **X**. $\mathbf{S}_X$ has to be determined by using an in-control part of **X**.

T^2 analyses each individual multivariate observation vector of a particular time instant k (1,...,n). Therefore, T^2 plot gives alarms at particular time instants when the process is out of control. T^2 values are distributed as F distribution and their upper control limit (UCL) is given by :

$$UCL_{T^2} = \left(\frac{p(n^2 - 1)}{n(n-p)} F_{\alpha,p,n-p} \right) \tag{3.9}$$

n is the number of observations and p is the number of variables. Significance level is designated by α. $F_{p,n-p}$ (α) is the upper (100 α)th percentile of the $F_{p,n-p}$ distribution. Points above the control limit represent potential special cause of variation (Johnson and Wichern, 1998).

T^2 plot also gives alarm in a case that the variables deviate from their mean but still observed as in-control in univariate charts individually. Because T^2 method analyzes all the variables simultaneously and produce one T^2 value which brings the collective effect of the variables for a certain time.

3.2.2.6. Squared Prediction Error (SPE) :

SPE chart is a monitoring tool, which is based on the error of the model constructed by a statistical technique such as PCA, PCR or PLSR. SPE determines the deviation from the model plane for each particular time instant (k) where T^2 defines the deviation from the mean value. The T^2 chart detects whether or not the variation of the scores is greater than that can be explained by common cause. Monitoring of process only by using T^2 technique based on scores is not sufficient. If a completely new type of special event occurs which was not present in the in-control data used to develop the model, then new observations will move off the model plane. Such new events can be detected by SPE (Kresta *et al.*, 1991). In terms of model errors **e**, the SPE statistics is :

$$\begin{aligned} SPE_k &= \mathbf{e}_k \mathbf{S}_e^{-1} \mathbf{e}_k' \\ \mathbf{e}_k &= \mathbf{x}_k - \hat{\mathbf{x}}_k \end{aligned} \tag{3.10}$$

SPE_k is a scalar value for each time instant k. Error component $\mathbf{e}_k$ at time k is a 1xp vector which is the difference between the actual ($\mathbf{x}_{1xp}$) and the predicted ($\hat{\mathbf{x}}_{1xp}$) values of the data matrix $\mathbf{X}_{nxp}$ for a particular observation time (k). Estimate of covariance matrix of **e**, which is $\mathbf{S}_{e(pxp)}$, should be attained from an in-control part of the data set. $\mathbf{SPE}_{nx1}$ can be defined as T^2 of the error vector. Thus, upper control limit of SPE plot is the same as the T^2 limit which is given by Equation 3.9.

3.2.2.7. Contribution Plots :

The main purpose of SPM is to provide an effective process control by determining out of control cases due to special cause of variation and preventing them to repeat. Multivariate monitoring tools, T^2 and SPE charts, detect deviations from normal operating conditions by combining the information coming from all process variables in a single statistics at each time instant. However, they do not reveal the responsible variable(s) for the out of control situation. In this step, contribution plots are able to examine which variables contribute to that particular out of control state (Westerhuis *et al.*, 2000). Contribution plots determine which variable or group of variables has/have contributed to the out of control signals of T^2 and SPE plots (Mason and Young, 2000). When an out of control alarm is received in T^2 and SPE plots at time k, contribution plots of T^2 and/or SPE statistics are constructed by calculating the contribution of each variable at that particular time instant.

Contribution of variables to T^2 is given as :

$$c_{k,a,j} = \frac{t_{k,a}}{\lambda_a} P_{j,a} X_{k,j}$$
$$C_{k,j} = \sum_{a=1}^{r} (c_{k,a,j}) \qquad (3.11)$$

c is a (nxrxp) matrix including contributions of r (1,..,a,..,r) latent variables on the p (1,..,j,..,p) original variables for a particular time instant (k)(1,..,k,..,n). **t** is a nx1 column vector and $t_{k,a}$ is the value of the ath score at time k. λ has the dimension of px1. λ_a is the ath eigenvalue obtained from covariance matrix of $\mathbf{X}_{nxp}$. **P** is the loading matrix which is pxr. $P_{j,a}$ is the loading of the ath score on the jth variable. $X_{k,j}$ is the value of the jth variable at time k. When calculating contribution, $c_{a,j}$ is set equal to zero if its sign is opposite to the value of the score $t_{k,a}$. If not, the contribution value remains the same. Then the total contribution ($C_{k,j}$) of p variables is obtained. **C** is a nxp matrix.

Contribution of variables to SPE is given as :

$$C_{k,j} = \frac{\left(\frac{e^2{}_{k,j}}{S_{e(j,j)}} \right)}{SPE_k} \qquad (3.12)$$

C is a (nxp) matrix defining contributions of p variables for each time instant (k). $e_{k,j}$ is the error value of the jth variable at time k. SPE_k is the squared prediction error value for the kth time instant. $S_{e(j,j)}$ is the variance value for the jth variable.

3.3. Cause and Effect Diagrams (Fishbone Diagrams)

The cause and effect diagram is a problem analysis tool that provides a systematic way of defining the problem and potential causes that create or contribute to the problem (Figure 3.5). This type of analysis attempts to identify the root causes for a problem. The lines coming off the core horizontal line are the main causes and the lines coming off those are sub causes. Cause and effect diagrams do not have a statistical basis, but are excellent aids for problem solving and quality improvement. Generally, the main categories of causes in a fishbone diagram are ;

- Methods, machines and materials
- People, places and procedures
- People, policies and surroundings
- Suppliers, system and skills

(Montgomery, 2000).

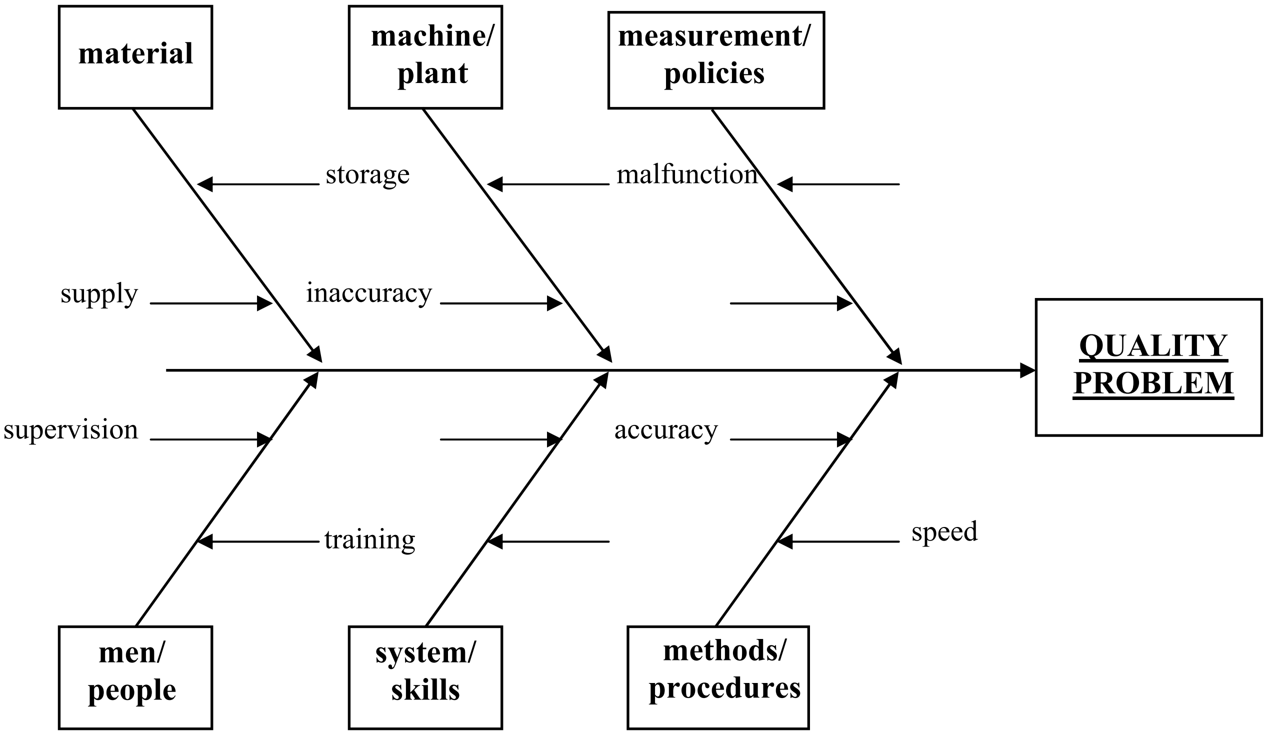


Figure 3.5. Sample Cause and Effect Diagram

3.4. HACCP and SPM Integration

HACCP system is based on the CCP (Critical Control Point) approach and one of the main steps of HACCP is monitoring of these CCP's. The fundamental philosophy behind both HACCP and SPM is prevention. Integrating SPM with HACCP system could further enhance this prevention philosophy by data analysis. Data analysis displays warnings for out of control states, and with appropriate corrective actions, failures can be prevented. The use of SPM techniques at the monitoring step of CCP's will be useful to capture the possible out of control cases. Therefore, SPM will provide a more effective HACCP system. The use of SPM in HACCP system provides effective monitoring and verification of the process and control of process variability. Also, a process within critical limits required in HACCP plan and quality of product in addition to safety are obtained by the combination of SPM techniques.

Statistical process monitoring is an integral part of HACCP system. The use of SPM techniques will provide a very comprehensive analysis of data, which are collected from a food production process within a HACCP system. SPM integration will improve the HACCP system and the safety of the produced food will be increased by eliminating potential risks. In addition to benefits for safer food production and public health, this integration will bring some economical benefits to the processing company. Increasingly, national and international legislation and trade requires the implementation of HACCP in food businesses (Jouve *et al.*, 1997). Therefore, companies with more effective HACCP systems will compete better with the other companies in the industry. Since international trade is stricter on food safety, the companies will have an advantage in international competition. HACCP system integrated with SPM will also remove the risk of the end product, thus, reducing the cost for destruction or reprocessing of the end product.

Food processes are multivariate in nature and the quality of a food product is a combination of several properties (Sahni *et al.*, 1999). Therefore, food production process includes many correlated variables. Although it is possible to control the process with univariate methods, multivariate methods provide a more comprehensive analysis of the data. Since process monitoring with data analysis is an important part of HACCP system, the food process within the system should be monitored with multivariate SPM techniques. However, the studies on HACCP system integrated with multivariate SPM techniques, which enables simultaneous analysis of the variables are limited in the literature.

3.5. Literature Survey

SPM has long been used in various fields including the food industry. In some past studies, simple univariate SPM methods such as histograms and individual charts were used in food production (Miller and Balch, 1991; Buco, 1990) and in HACCP systems.

In a study carried out by Hayes *et al.* in 1997, two univariate SPM methods are used in the filling section of a milk plant within HACCP system. Relative Light Units data displaying the hygiene status of the product were collected over 3 months from a milk-filling machine which is a CCP and were analyzed using CUSUM and Individual charts. The SPM techniques used in the study gave different patterns for the process; however, both displayed clear warnings for a severe out of control situation. If some corrective actions had been taken before this point, the out of control case would have been prevented.

In another study, microbial data were collected from Butterball Turkey Company during a year and analyzed by using histograms and individual moving range control charts (Surak *et al.*, 1998). The data were *E.coli* counts collected from carcasses at the end of the chill system. The SPM tools showed that the process was stable and met specifications during the first part of the year. After a major change was made in the processing method, *E.coli* levels of the carcasses increased. The process was still capable but had lost its stability. This change was easily captured by histogram and control chart and some corrective actions were taken such as additional chlorinated spray cabinets, improvement of the first and final wash cabinets, chlorine levels check, increased water pressure and additional inspectors at some points on the process line. As a result of these corrective actions, the *E.coli* level was reduced and the process was again stable and capable during the third part. In addition, the average level of *E.coli* was lower than that of the first part. This pattern was also displayed successfully with the SPM tools.

Gonzalez-Miret *et al.* (2001) investigated validation of HACCP parameters with univariate and multivariate statistics in a poultry meat production plant. The parameters were microbial counts such as Total Count (TC), *Pseudomonas* (PS), *Enterobacteriaceae* (EB) and *Staphylococcus aureus* (SA) from two different control points: refrigeration and cutup/packaging. Samples were taken at three different stages for both control points. Pair comparison by Bonferroni Method and Analysis of Variance (ANOVA) were applied to control points and the differences of microbial loads at each stage were tested. According to the results significant differences were observed between stages especially in TC and EB counts and it was concluded that the decontamination effect of refrigeration is higher than the recontamination effect of handling.

In a more recent study Srikaeo and Hourigan (2002) considered the use of statistical process monitoring to enhance the validation of CCP's in shell egg washing process of an existing HACCP system. Their aim was to utilize SPM tools to evaluate whether the process is under control and to determine the capability of the process. Shell egg washing process is a CCP since *Salmonellae* may previously have contaminated the eggs and should be eliminated or reduced. The parameters were pH of wash water, temperature of wash water, temperature of rinse water and chlorine level. Univariate control charts were performed to the data of these parameters over a period of six months. The results showed that the control measures are satisfactory in terms of safe food production and capability studies indicated that all control measures are capable to their critical limits except chlorine level. According to these findings authors concluded that the process of shell egg washing process is well-designed in terms of food safety purposes and SPM enhance the validation of the HACCP system.

The literature involves many studies about the theory of multivariate SPM and its applications in various fields especially in chemical industry (Kresta *et al.*, 1991; MacGregor *et al.*, 1994; Kourti and MacGregor, 1995; Nijhuis *et al.*, 1997; Wikström *et al.*, 1998; Martin *et al.*, 1999; Conlin *et al.*, 2000). Multivariate statistical techniques have been studied for food industry. Buco (1990) proposed the use of multivariate statistical techniques in food production. Negiz *et al.* (1998) modeled a pasteurization unit empirically and studied T^2 technique in the system to determine the abnormal behaviors. Sahni *et al.* (1999) suggested the application of multivariate statistical analysis and design of experiment in product development in the food industry. Kösebalaban and Çınar (2001) also performed empirical modeling of a pasteurization unit and apply T^2 chart, contribution plots and parity space technique in order to decide abnormal situations and their causes. In the literature, Multiple Linear Regression (MLR) and Principal Component Regression (PCR) were generally used for nutritional and instrumental analysis studies respectively.

The use of multivariate statistical techniques in HACCP system was mentioned in 1999 by Çınar *et al.* Their study included cooked sausage processes within a HACCP system. A model describing the general behavior of the sausages temperature at a CCP was developed by PCA and T^2 and SPE techniques were performed. Univariate techniques such as X-bar and S charts were also used to provide additional information for diagnosis. The results indicated that the CCP can be more effectively monitored using multivariate statistical techniques rather than the univariate methods.

However, it is not encountered to any other published studies, which use multivariate SPM techniques in a HACCP program in the literature.

Chapter 4
MATERIALS AND METHODS

4.1. Material

The study was carried out in cooperation with a frozen food processing plant located in İzmir. The plant utilizes HACCP system as a safety management programme. One of the most microbiologically problematic productions of the plant is frozen red pepper production since pH of red pepper is between 5.0-6.0. Therefore, the specific process line chosen for the study was frozen (blanched) red pepper production and the flow diagram of this process is shown in Figure 4.1.

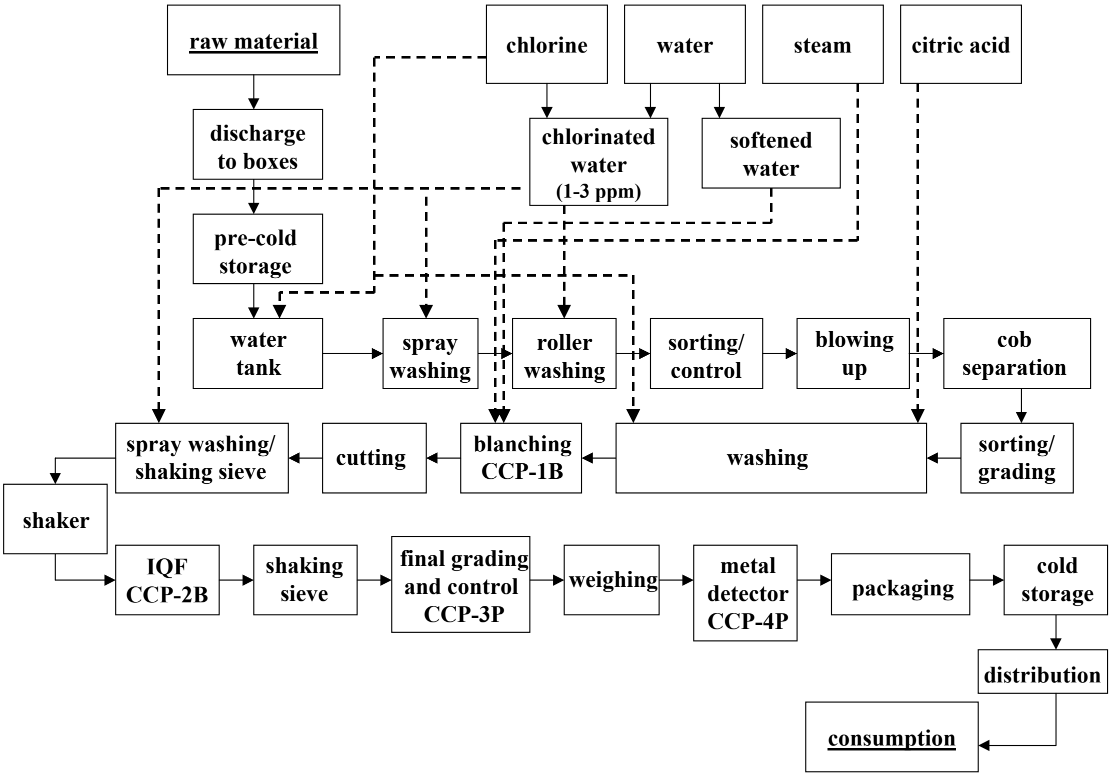


Figure 4.1. Flow Diagram of Frozen (Blanched) Red Pepper Production

The material was the data, which were collected from the frozen red pepper production and multivariate SPM techniques were applied to these data in this study. All the data were obtained from HACCP documentation.

4.2. Process

The raw material is discharged into boxes and stored in a cold storehouse prior to the process. Pre-cold storage enables to control the increase of microorganisms. During the process, the raw material is cleaned at several steps. The first step is water tank, which enables to eliminate relatively heavy foreign materials such as stones. Chlorine is added into the water in the tank to decrease the initial microbial load. Chlorinated water is also used at the other steps of cleaning which are spray and roller washing. After washing, the foreign materials, undesirable parts of the material and improper units are removed. Then the material is washed again. Chlorine and citric acid are used at this step as a second barrier for the microorganisms. The first CCP (CCP-1) is blanching step. The aim of this heat treatment is enzyme inactivation. The material is blanched at around 90-95^{o}C for 70 seconds. The hazard of concern at this step is biological. After blanching, the material is cut into the desired size. Since the cutting step may increase the microbial load, spray washing is applied by means of chlorinated water. Then the material is sieved and the washing water is removed by the shakers. The second CCP (CCP-2) is Individual Quick Freezer (IQF). Corresponding hazard at the freezing step is biological. The material is very rapidly frozen not to damage the structure of the vegetable at this step. Later, another shaker is utilized and the crystals are removed from the surface. Final grading and control which is the third CCP (CCP-3) provides the last inspection of the material where the hazard is physical. Then the material is weighed, passed through the metal detector and packaged. Metal detection is the fourth CCP (CCP-4) where the corresponding hazard is physical. As the last step, the packaged material is sent to cold storage (-18^{o}C) and distributed for retail sale and consumption. The CCP's of the process and their hazards of concern are shown at Table 4.1.

Table 4.1. Concerning Hazards of CCP's

CCP-1B	Inability of decreasing pathogen microorganism load due to insufficient thermal treatment
CCP-2B	Growth of pathogen microorganisms due to insufficient freezing
CCP-3P	Presence of foreign materials such as stone, glass, etc.
CCP-4P	Presence of metals

4.3. Important Parameters of the Process

The important parameters of the process, which should be carefully detected during the process to obtain a high quality and safe product, are given in this section.

4.3.1. Raw Material

High quality raw material is the most important priority for a high quality product. Amount of foreign material and rotten, burst, diseased or crumpled units in the raw material are the quality criteria. pH of red pepper is generally between 5.0-6.0. Due to this low acidity (near to neutral), it is susceptible to microbial spoilage. Brix, which is the measure of total soluble solids, is around 6.5-8.5 for red pepper raw material. Therefore, besides the quality of the raw material, its physical properties are quite important.

4.3.2. Chlorine and Citric Acid Content

Chlorine and citric acid content of washing water is an important parameter since washing with chlorinated and acidified water is the first hurdle for microbial growth during the process. Chlorine is added in sodium hypochlorite form at 15 to 20 ppm. The effect of chlorinated water is much greater at pH 2.5. In order to obtain this synergic effect, pH of water is decreased by using citric acid.

4.3.3. Blanching Time and Temperature

The aim of blanching is to inactivate enzymes, which causes undesirable colors and flavors that may develop during processing and storage. Peroxidase is the indicator enzyme of inactivation. The blanching parameters on the average are between 90-95^{o}C approximately 70s. In HACCP procedure, blanching should be applied at 85^{o}C for 55 seconds minimum. Since these parameters also provide pasteurization, blanching is accepted as a second barrier for the microorganisms in the process after chlorinated and acidified water treatment.

4.3.4. Individual Quick Freezer (IQF)

The product should be at minimum -12^{o}C at the outlet of the IQF. Defrost status and the condition of IQF honeycombs influence the efficiency of the IQF. It affects the microbial quality of the product.

4.3.5. Final Grading and Control

Final grading and control is applied by the personnel at the exit of the IQF and before packaging. This is the last point to control the foreign material existence in the product. These foreign materials can be plant originating, stones, cord, glass, hair, plastic, etc., which may arise from raw material, equipment, personnel, packaging materials, plant and surroundings.

4.3.6. Microbial Counts

Microbial analyses such as Total Viable Count, Total coliform, *Escherichia coli*, yeast, mold, *Staphylococcus aureus*, *Enterobacter* and *Listeria monocytogenes* are performed to the sample which is taken from the line before packaging and the results are evaluated as the verification of CCP-1 (blanching) and CCP-2 (IQF).

4.3.7. Metal Detector

Presence of metal in the product is a very considerable physical risk. Therefore, the efficiency of the metal detector and its calibration is observed regularly. Metal particles such as screw, nail, watch piece etc. can originate from equipment, personnel, etc.

4.4. Variables

In order to apply multivariate SPM techniques, multivariate data including raw material, process and end product measurements were collected from various process steps including CCP's during the production. The variables, which were used in data analysis, are given in Table 4.2.

Table 4.2. Variables of the Process

Raw Material Properties	Plant origin foreign material Rotten Burst Diseased Crumpled Brix	x1 x2 x3 x4 x5 x6
Process Conditions	Blanching time Blanching temperature	x7 x8
End Product Properties	Foreign material Microbial counts Total Viable Count (TVC) *E. coli* Yeast Mold	y1 y2 y3 y4 y5

Since the quality of raw material is the most important criterion for the product quality, raw material properties (x1 to x6) were included in the study. Blanching time & temperature (x7 and x8) and foreign material (y1) were the variables collected from CCP-1B and CCP-3P, respectively. Finally, microbial counts (y2 to y5) which are the verification of a safe production were selected as variables. Temperature measurements of the IQF are not collected by the plant since it is computerized to the instrument and no considerable deviation is expected. Metal detector results are attributes type of data. Observations are recorded as conforming/nonconforming or yes/no in attributes data. Therefore, metal detector observations are not in the same form with the other variables. IQF and metal detector data were not included in the data set because of above reasons even they are the data of a CCP.

The data collected from the frozen red pepper production were extracted by eliminating the unmatched measurement sets for a particular time instant and obtained data set including 13 variables and 147 observations were used in the modeling and monitoring the process. A set of 24 observations within the data set (observations [89-112]) was selected as in-control set (Appendix B) since it describes normal operating conditions. Descriptive statistics of in-control data set is given by Appendix C. The data and the in-control data sets were pretreated prior to the model development. Principal Component Analysis (PCA), Multiple Linear Regression (MLR), Principle Component Regression (PCR) and Partial Least Square Regression (PLSR) are the multivariate methods which were employed for data based modeling of the process in this study.

Hotelling's T^2 Plot, Squared Prediction Error (SPE) Chart and Contribution Plots are the multivariate tools used to monitor the process pattern visually. Figure 4.2 illustrates the procedures which were followed.

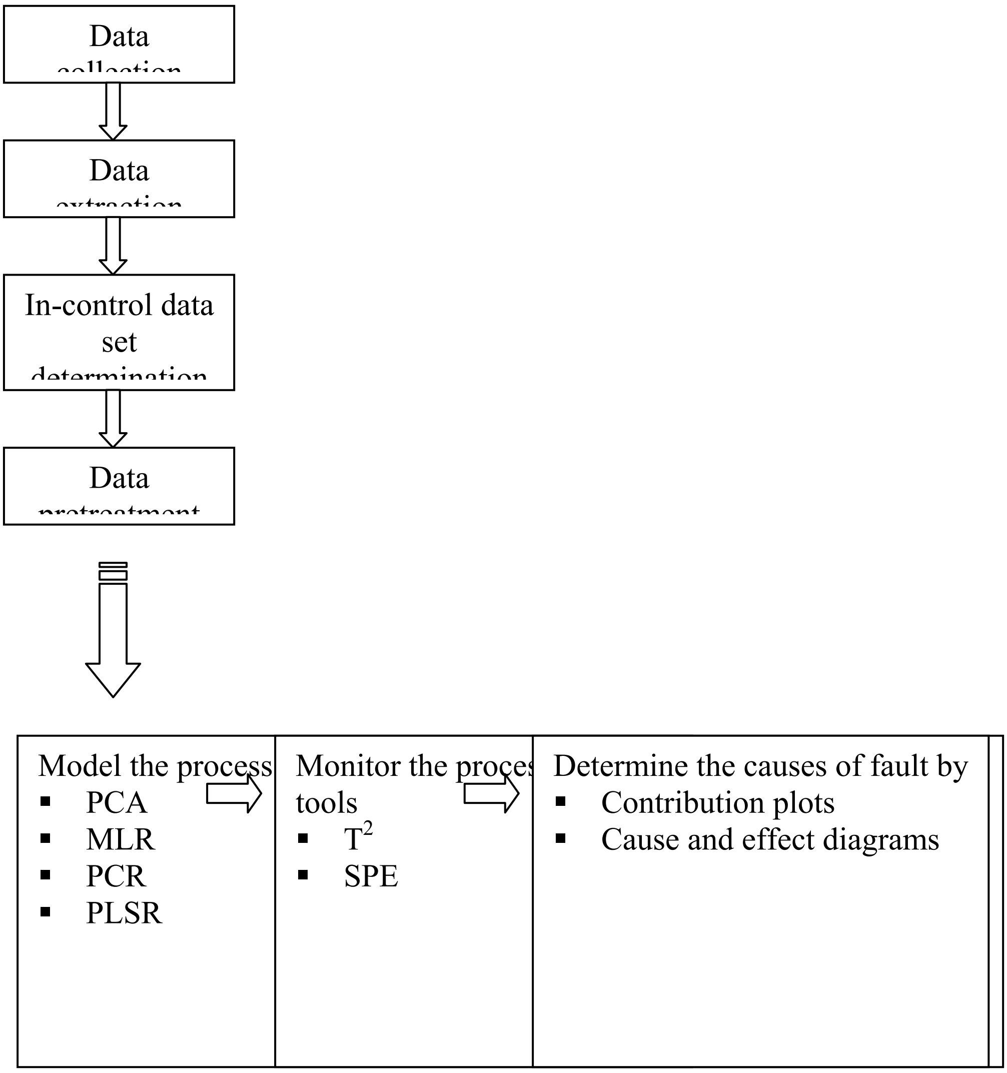

Data collection
Data extraction
In-control data set determination
Data pretreatment
Model the process
PCA
MLR
PCR
PLSR
Monitor the process tools
T2
SPE
Determine the causes of fault by
Contribution plots
Cause and effect diagrams

Figure 4.2. The Monitoring Procedure Followed in the Study

4.5. Data Pretreatment

Pretreatment of the data in statistical analysis is often very important. It provides transformation of the data into the most suitable form for statistical analysis. Transformation is performed when the variables are measured in different units or as a technique to remove noise (Sahni *et al.*, 1999).

4.5.1. Transformation:

When the data is non-normally distributed, transformation is applied to match a normal distribution. In the literature, logarithmic transformation was used for gamma type of distribution, square root transformation for Poisson type of distribution (Schaffner, 1998) and power (0.25) transformation for the data including zero values (Wold *et al*., 2001). Table 4.3 shows several types of transformations.

Table 4.3. Transformations

Original data	Transformation	
Normal	No transformation	y
Gamma	Log	log (y)
Poisson	Square root	$y^{0.5}$
Zero values	Power	$y^{0.25}$

The data sets of variables used in this study indicate different properties, such as microbial measurements or zero values. Therefore, different transformations were tried for each variable. Within these transformations, the one, which gave the best result, was chosen for that particular variable to use in transformation to normality. The best results were obtained when no transformation was applied to **X** data set (raw material properties) and 0.25 power transformation was applied to **Y** data set (end product properties).

Normality plot (or Q-Q plot) is used to assess if the observations come from a normal population or they violate the normality assumption. The procedure for the construction of a normality plot includes 3 steps :

- Order the original observations $X_1,X_2,...,X_n$ to get $X_{(1)},X_{(2)},...,X_{(n)}$ and their corresponding probability values (1-1/2)/n, (2-1/2)/n, ..., (n-1/2)/n
- Calculate the standard normal quantiles $q_{(1)},q_{(2)},...,q_{(n)}$
- Plot the pairs of ordered observations and quantiles and examine the straightness of the curve (Johnson and Wichern, 1998).

In the study, normality of the transformed data was also checked by normality plots in order to confirm that a normally distributed data set was obtained with transformation procedure.

4.5.2. Scaling:

Before the model is formed, it is usual to scale the data to zero mean and unit standard deviation in order to obtain variables with the same dimension (Sahni *et al.*, 1999; Runger and Montgomery, 1997). Scaling (autoscaling or standardization) includes mean centering and variance scaling. The mean value of each variable is calculated from the in-control set and subtracted from corresponding variable to achieve the mean centered values. When the variables in a block are measured in different units, variance scaling is also used. Each variable is divided by the corresponding standard deviation that is obtained from the in-control set, so that the variance of every variable is unity (Geladi and Kowalski, 1986). Figure 4.3 displays the illustration of mean centering and variance scaling.

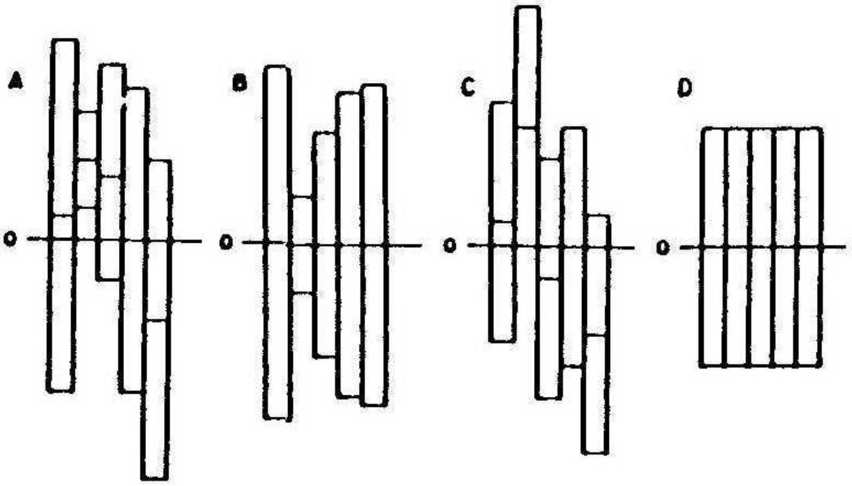


Figure 4.3. Data Pretreatment
The data for each variable are represented by a variance bar and its center. (A) Most raw data look like this. (B) The result after mean centering only. (C) The result after variance centering only. (D) The result after mean centering and variance scaling. (Geladi and Kowalski, 1986).

In the study, scaling was applied to the entire data to provide the dimension convenience among the variables.

The multivariate SPM methods were applied to the data of some certain process steps and the results were displayed visually in charts. The software that was used for calculation and charting procedures was MATLAB. The following gains were expected;

- to construct a process model with independent latent variables which were obtained by PCA, PCR and PLSR methods

- to model the relation between process data with the end product properties by regression methods; MLR, PCR and PLSR
- to capture the out of control points by monitoring Hotelling's T^2 and SPE charts of the latent variables
- to assess the precision of the constructed models by SPE method
- to determine the contributing variables of the out of control points by contribution plots and cause & effect diagrams

4.6. Multivariate SPM Methods and Tools

All the multivariate models were built with the in-control data set. The procedures employed for the multivariate methods and tools in this study are explained below :

PCA analysis was used to model the process variables data **X** which is composed of raw material properties and process conditions (x1 to x8) (Figure 4.4-a) and quality variables data including end product properties (y1 to y5) (Figure 4.4-b) individually. The PCA models were constructed by using the in-control sets and applied to the entire data. No transformation was used in the former data matrix however power transformation was used for the latter.

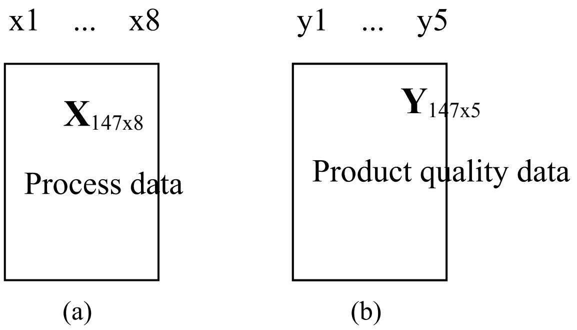


Figure 4.4. X and **Y** Matrices Used in the PCA Method

In the PCA analyses of process data and product quality data matrices, normality plots of in-control error vectors were constructed in order to examine their distribution. All of the plots constructed by using in-control set were intended to check the model. Score plots (**t1** vs **t2**, **t1** vs **t3** and **t2** vs **t3**) with 95% control ellipses were also investigated. These plots were helpful to evaluate the distribution of the score values. 100(1-α)% control ellipse is given by all score vectors for a particular time instant (k) satisfying Equation (4.1). $\mathbf{S}_t$ is the p by p covariance matrix of in-control scores. $\chi^2_{(\alpha,p)}$ is the upper α percentage point of the chi-square distribution with p degrees of freedom.

$$\mathbf{t}\mathbf{S}_t^{-1}\mathbf{t'} \leq \chi^2_{(\alpha,p)} \tag{4.1}$$

Hotelling's T^2 plot does not give good results if the variables, which are directly used in the method, are correlated. T^2 can also be computed from the scores of PCA, PCR and PLSR, which are uncorrelated (Miller *et al.*, 1998). Thus, T^2 plot was applied to the score matrix (**T**) of PCA model in order to capture the out of control status due to the deviation from the mean values. Similar to the control ellipses, 95% confidence level or 0.05 significance level (α=0.05) was used for T^2 limit. SPE chart was employed to determine the out of control measurements resulting from nonconforming to the model. In addition, contribution plots were applied to decide the contributing variables to the out of control points of T^2 and SPE plots constructed.

The relation between **X** and **Y** was examined by regression methods. Raw material properties and process conditions (x1 to x8) were used as **X** matrix and end product properties (y1 to y5) were used as **Y** matrix. The models were formed by using the in-control data set including 24 observations, 8 process variables (x1,..,x8) and 5 quality variables (y1,..,y5). Transformation was not applied to the **X** matrix however power transformation was applied to the **Y** matrix to approximate it to normal distribution. Then, the regression models were applied to the complete data sets $\mathbf{X}_{147x8}$ and $\mathbf{Y}_{147x5}$ to compute model residuals (Figure 4.5).

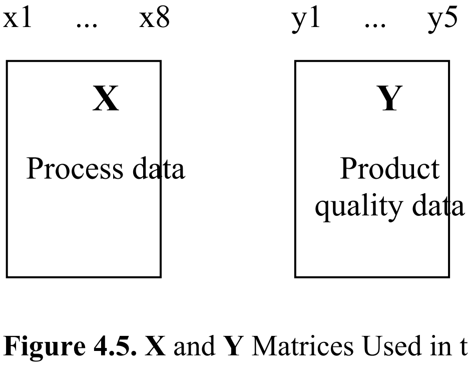


Figure 4.5. **X** and **Y** Matrices Used in the Regression Methods

In the MLR method, normality plots of in-control errors of **Y** were investigated and estimation of in-control errors were plotted against estimation of **Y** values. The aim was to check the accuracy of the regression model.

The first step of PCR analysis is to obtain principal components of **X** ($\mathbf{T}_{147x4}$) as in PCA method. Therefore the plots belong to **X** matrix which are normality plots of in-control error of **X**, score plots, T^2 plot of score matrix, SPE plot of **X** and contribution plots of T^2 out of control points would be completely the same as in PCA analysis. These plots were not represented again in the PCR method because of this reason. The two different one were the plots of **Y** which were normality plots of in-control error of **Y** and SPE plot of **Y**. Estimated in-control errors were also plotted against estimated **Y** values for each variable to analyze the regression model.

A similar pathway as in the PCA method was followed in PLSR analysis. Normality plots of in-control errors of **X** and **Y** were employed. Scatter plots of scores of **X** and **Y** matrices were plotted. Hotelling's T^2 plots were utilized for two score matrices ($\mathbf{T}_{147x4}$ and $\mathbf{U}_{147x4}$) of PLSR model. SPE charts were constructed to check the accuracy of the models of **X** and **Y**. The plots of **Y** values versus estimated in-control errors were also practiced. 95% confidence level (α=0.05) was used for control ellipses of score plots and T^2 limit.

Cause and effect diagrams were employed for all CCP's and the possible causes were listed for potential hazards at these CCP's. Experiences of the plant on their HACCP program and the major problems, which were documented, were based in the construction of cause and effect diagrams.

Chapter 5
RESULTS AND DISCUSSION

5.1. Data Pretreatment

In this study, transformation and scaling were applied prior to the multivariate analysis in order to achieve normal distribution, zero mean and unit standard deviation. Transformation and scaling, which were applied to the data matrices, are displayed in Table 5.1. No transformation was applied to the **X** matrix since normality plots of **X** variables displayed patterns, which were very close to the normal distribution. However, the normality plots of **Y** variables showed non-normally distributed patterns. Thus, **Y** matrix was pretreated with 0.25 power transformation to approximate normality.

Table 5.1. Pretreatment Applied to the Data

Data matrix	Transformation	Scaling
X	Not used	Scaled
Y	0.25 power transformation	Scaled

5.2. Principle Component Analysis (PCA)

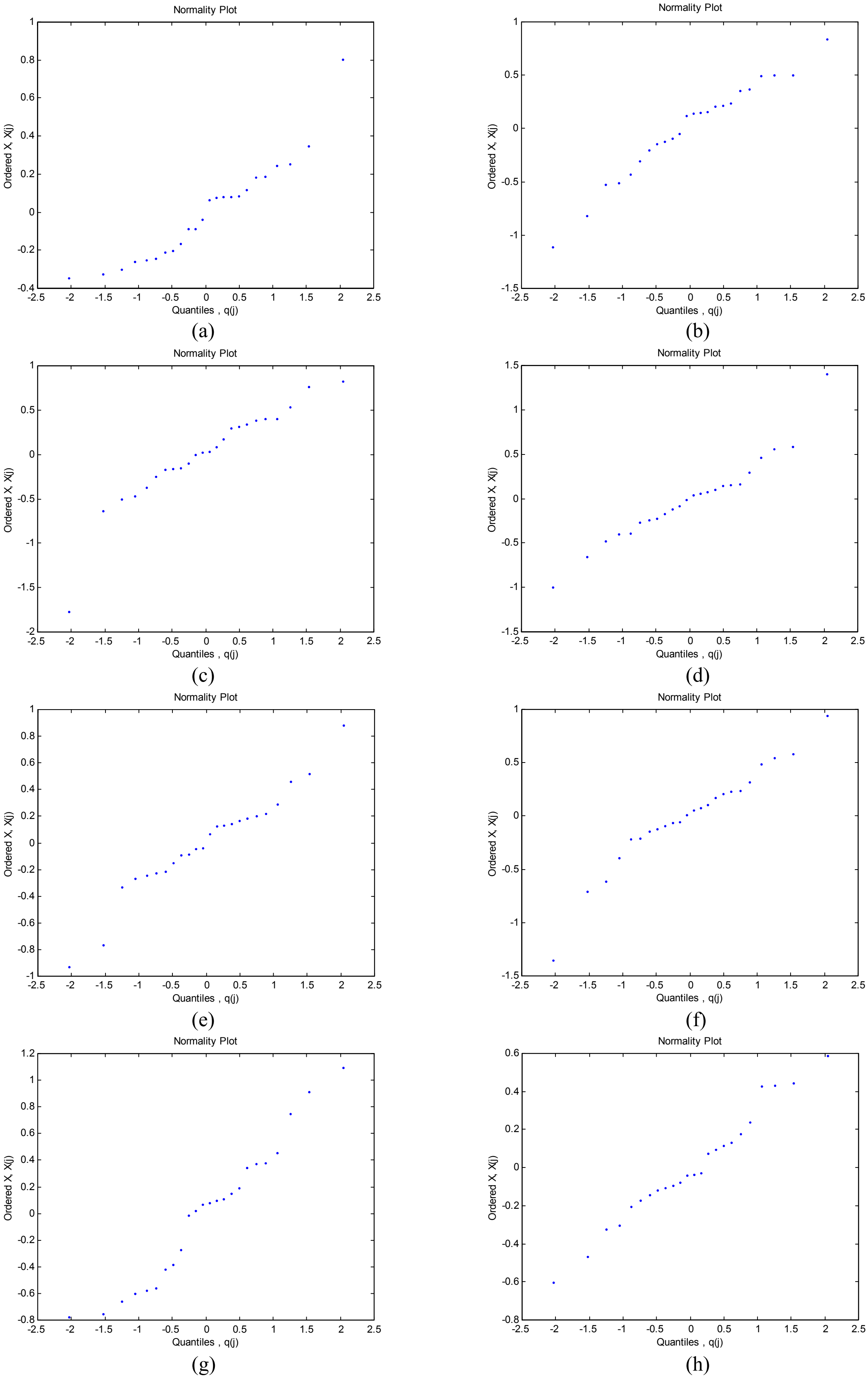
X matrix including raw material properties and process conditions was analyzed by means of PCA method. The variables of the **X** matrix are shown in Table 5.2.

Table 5.2. Variables of **X** Matrix

Raw Material Properties	Plant origin foreign material Rotten Burst Diseased Crumpled Brix	x1 x2 x3 x4 x5 x6
Process Conditions	Blanching time Blanching temperature	x7 x8

In PCA method, the model was constructed by using the in-control data set. No transformation was used and 4 principal components explaining 80.99% of the variance of **X** data were selected for PCA analysis. The first principal component accounts for 40.21% of the total variation in data set, whereas the second, third and forth components account for 15.84%, 15.19% and 9.75% of the variance.

Normality plots of the in-control errors are shown in Figure 5.1 to evaluate the model development. If the model fully explains the data structure, in-control errors should be normally distributed. The normality plot should represent a straight line if the values are normally distributed. In respect to Figure 5.1, distribution of the in-control errors are near to normality. Thus, it is possible to say that the model successfully explains the **X** data.



(a)

(b)

(c)

(d)

(e)

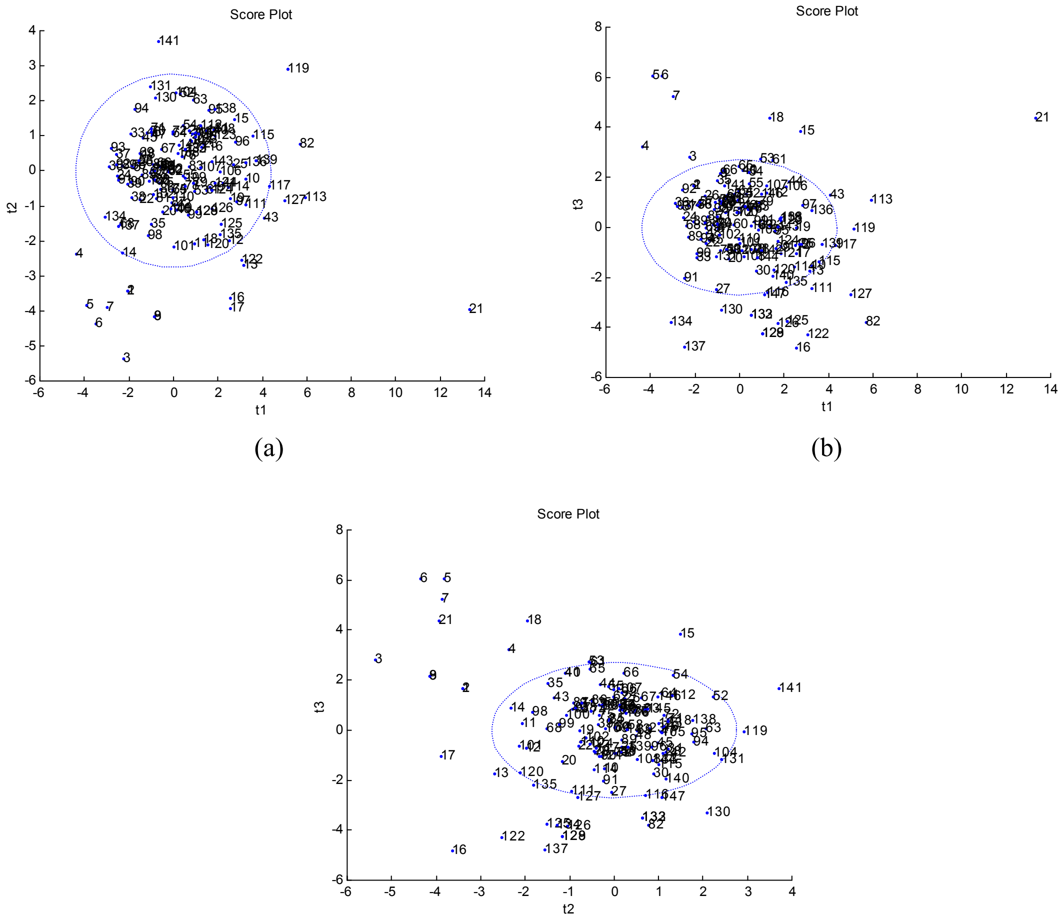
(f)

(g)

(h)

Figure 5.1. Normality Plots of In-control Errors in PCA Analysis of **X** Data. Figure 5.1-a to Figure 5.1-h represent the normality of errors of variables 1 to 8, respectively.

Figure 5.2 illustrates the scatter diagrams of PCA scores (**t1-t2, t1-t3** and **t2-t3**) and their 95% control ellipse. According to Figure 5.2-a (**t1-t2** plot), observations 1, 2, 3, 4, 5, 6, 7, 8, 9, 13, 16, 17, 21, 43, 82, 113, 119, 122, 127 and 141 are expected to be out of control since they are outside of the 95% control ellipse in which the observations are highly concentrated. Similarly, observations 1, 2, 3, 4, 5, 6, 7, 8, 9, 13, 14, 15, 16, 17, 18, 21, 43, 82, 111, 113, 119, 122, 125, 126, 127, 128, 129, 130, 132, 133, 134, 137 and 141 are expected to be out of control as depicted in Figures 5.2-b and c (**t1-t3** and **t2-t3** plots).



(a)

(b)

(c)

Figure 5.2. Score Plots in PCA Analysis of **X** Data. Figure 5.2-a to Figure 5.2-c represent **t1** vs **t2**, **t1** vs **t3** and **t2** vs **t3**, respectively.

The superiority of T^2 plot of scores to scatter plot of scores is its ability to consider all of the scores simultaneously and keep the time sequence of observations. Figure 5.3 is the T^2 plot of **X** scores obtained from PCA analysis. Numbered observations are out of control points since they exceed the T^2 control limit with 95%

confidence level. T^2 values which gives deviation from the mean, is considerably high at the beginning and at the end of the frozen red pepper production season. Out of control points of T^2 plot exactly match with the observations which were expected to be out of control in score plots (Figure 5.2).

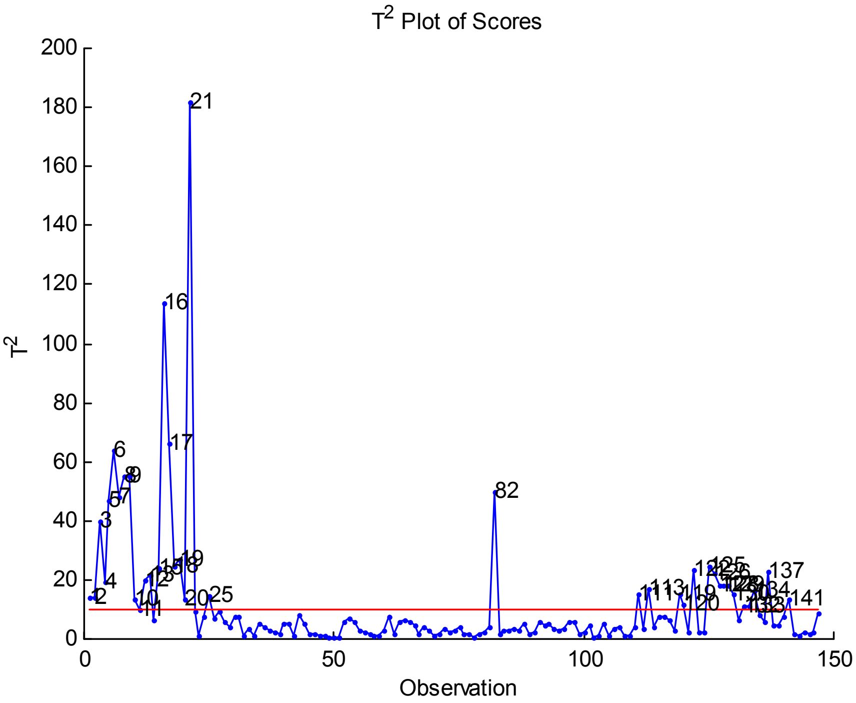
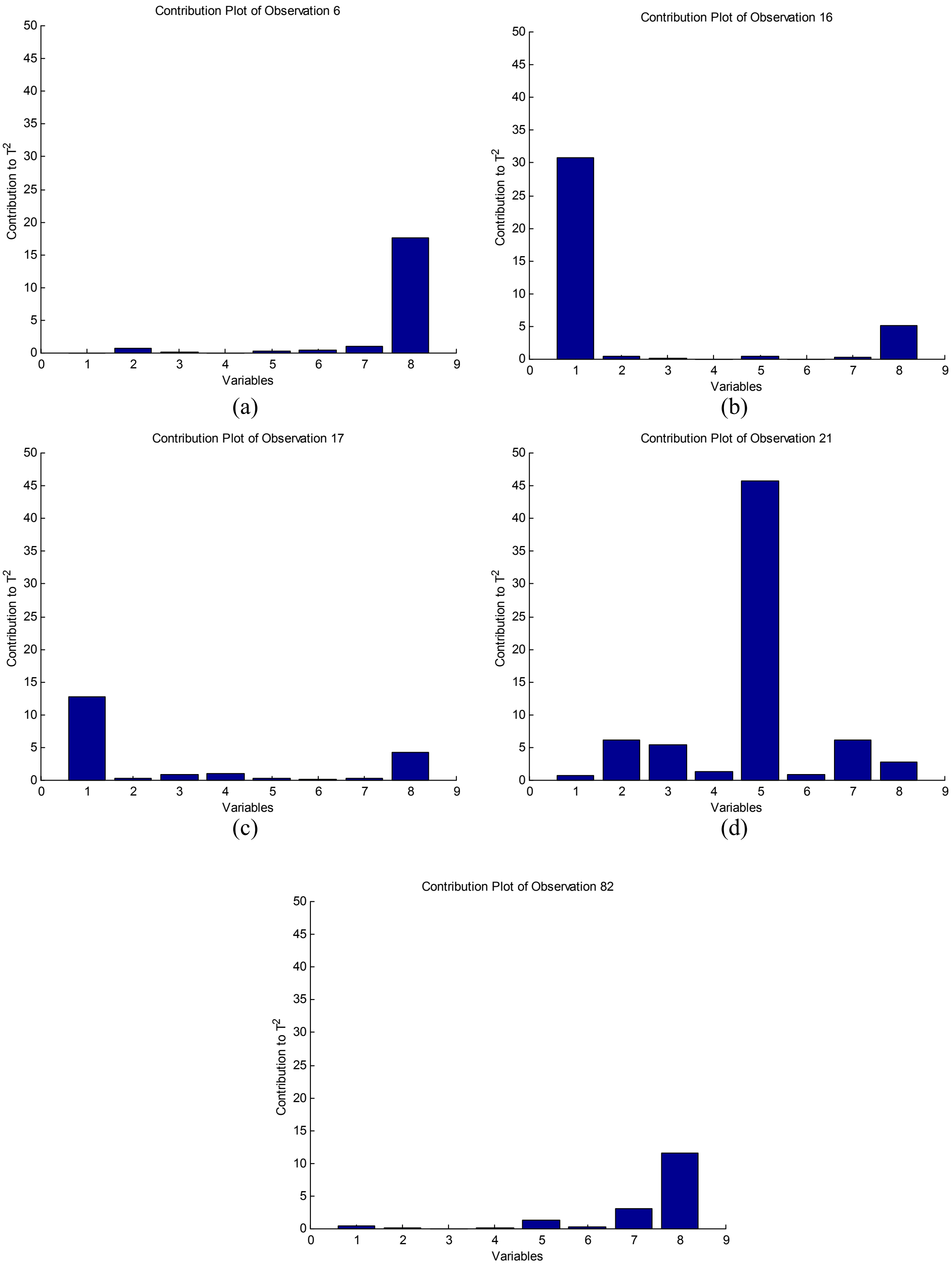


Figure 5.3. T^2 Plot of Scores in PCA Analysis of **X** Data

Observations 6, 16, 17, 21 and 82 are the most problematic ones among out of control points. Contribution plots of these observations are shown in Figure 5.4. In respect of Figure 5.4-a, blanching temperature (x8) seems to be the main factor, which causes the out of control state for observation 6. x8 has its lowest degree, which is 85°C, in observation 6. This temperature is also the minimum requirement for the blanching procedure of the HACCP plan. Contributions of plant origin foreign material (x1) and blanching temperature (x8) are the important factors both for observation 16 and 17 (Figure 5.4-b and c). The plant origin foreign material has the highest two values in these observations and the blanching temperature is 89°C. Crumpled raw material (x5) mostly contributes to observation 21 since it reaches its highest value in the entire data (Figure 5.4-d). Contributing variables to observation 82 are blanching temperature (x8) and blanching time (x7) (Figure 5.4-e). Blanching temperature is 98°C with its highest value at this point and blanching time is 50s. This time is very close to the minimum requirement for blanching procedure in HACCP plan.



(a) (b)

(c) (d)

(e)

Figure 5.4. Examples of T^2 Contribution Plots in PCA Analysis of **X** Data. Figure 5.4-a to Figure 5.4-e represent the contribution plots of observations 6, 16, 17, 21 and 82, respectively.

Contribution percentages of variables were determined by analyzing all of the T^2 out of control points for their contribution plots. Contribution of **X** variables to T^2 is in Figure 5.5. According to the figure, the most important contributing factors are plant origin foreign material (x1) with 38% contribution and blanching temperature (x8) with 32% contribution.

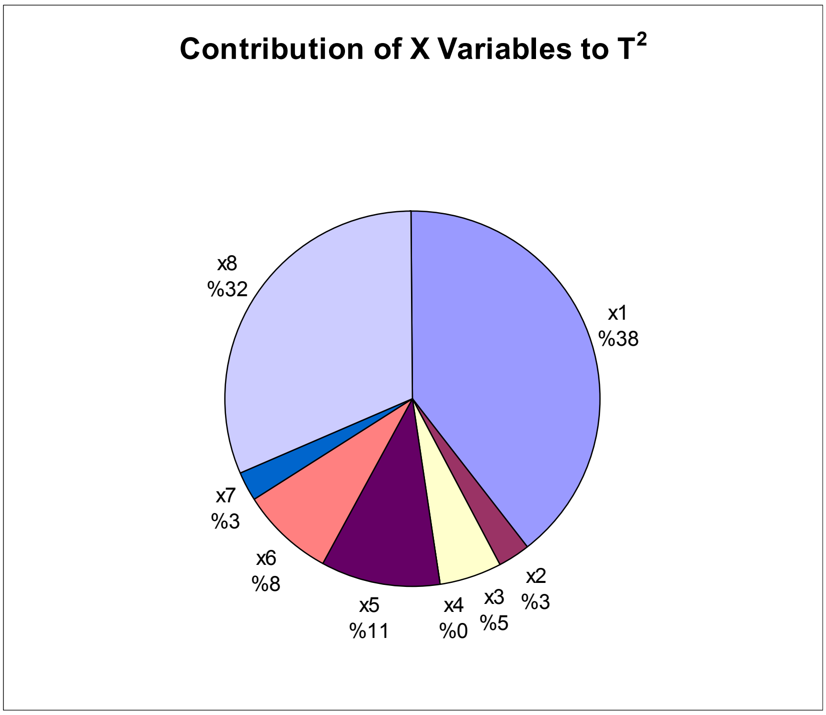


Figure 5.5. Contribution of **X** Variables to T^2

In SPE calculations of PCA model of **X** data, negative values in $\mathbf{eS_e^{-1}e'}$ computation (Equation 3.10) were observed. In theory, the covariance matrix of errors, $\mathbf{S}_{e(pxp)}$ has to be positive definite matrix. In other words, it has non-negative eigenvalues so that the multiplication of $\mathbf{e_k S_e^{-1} e_k'}$ are assured to be positive at all times. If a matrix is not positive definite, its rank is less than its dimension (singular matrix). For $\mathbf{S}_e$, which is a pxp square matrix, its rank is less than p in this PCA analysis. Its columns or rows are not linearly independent.

When the number of principal components retained in the model (r) is increased to 8, the singularity problem of covariance matrix has been solved. The rank of $\mathbf{S}_e$ was determined as 8 (=p) whereas it was 4 in the previous PCA model (r=4). The SPE calculation did not produce non-negative values. The chart is given in Figure 5.6. All of the observations are very close to the limit in SPE plot and many of them exceed the limit. However, the general pattern and the out of control points of SPE plot are quite similar to those of T^2 plot. Some observations such as 43, 44 and 45 are determined as out of control in SPE plot but not in T^2 plot, because these observations deviates from the model not from the mean value.

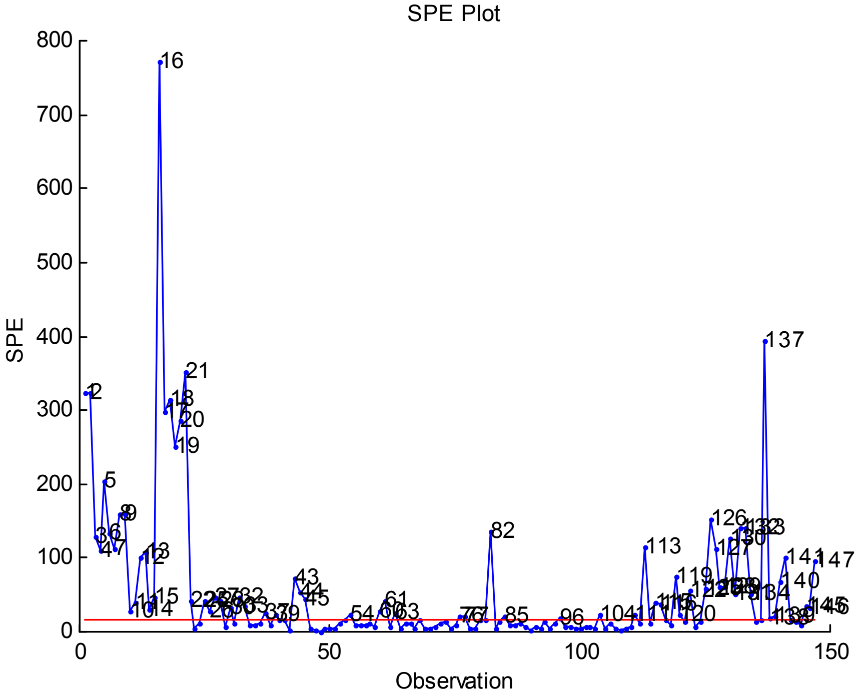


Figure 5.6. SPE Plot in PCA Analysis of **X** Data

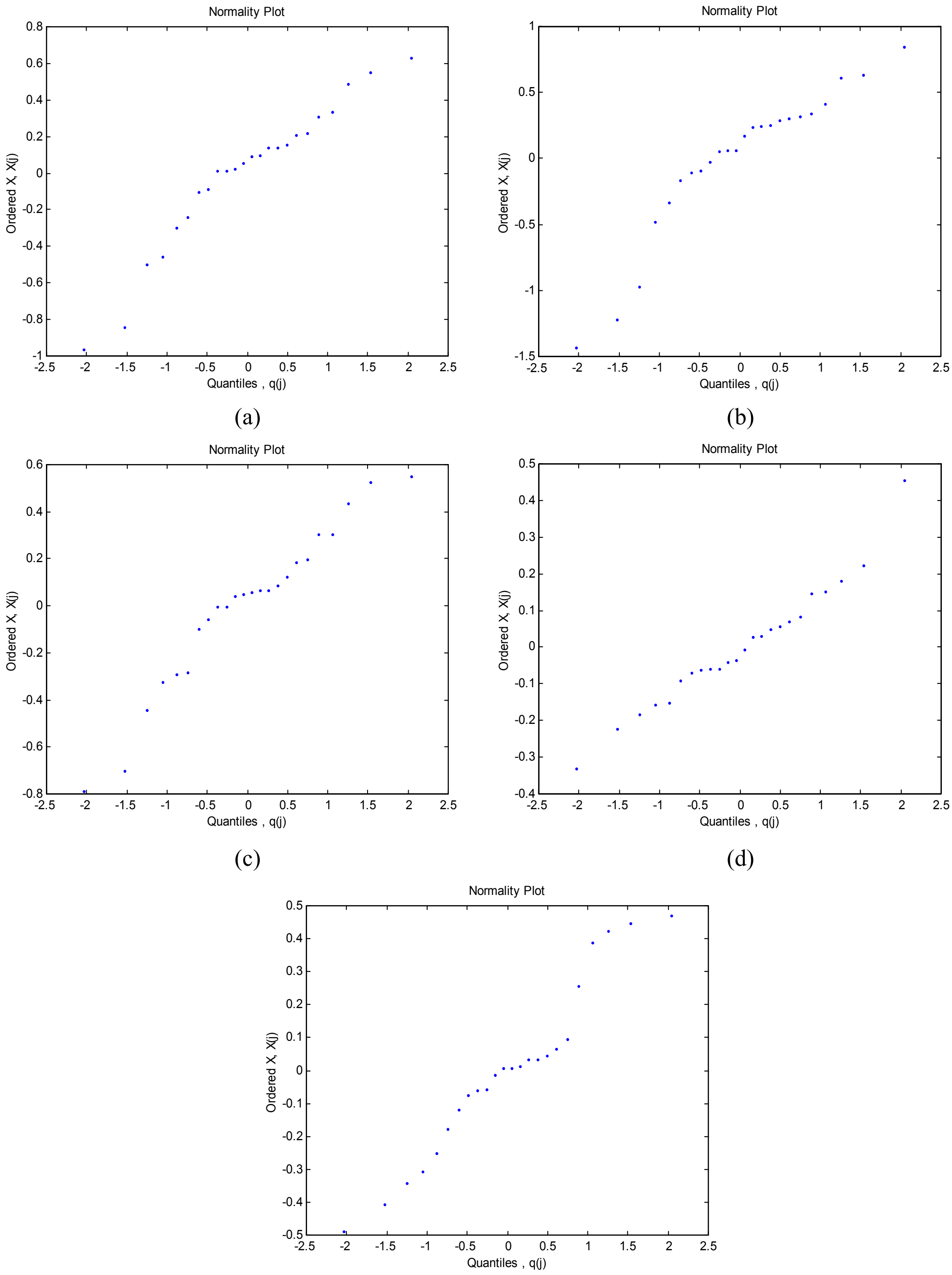
Y matrix including end product properties was also analyzed by means of PCA method since foreign material is the parameter of CCP-3 and microbial counts are the verification of CCP-1, which is blanching. The variables of the **Y** matrix are shown in Table 5.3.

Table 5.3. Variables of **Y** Matrix

End Product Properties	Foreign material	y1
	Microbial counts	
	Total Viable Count (TVC)	y2
	E. coli	y3
	Yeast	y4
	Mold	y5

0.25 power transformation was applied to the data matrix and 3 principal components explaining 86.44% of the variance of **Y** data were selected for PCA analysis. The first principal component explains 40.10% of the variability, whereas the second and third account for 29.35% and 16.99%.

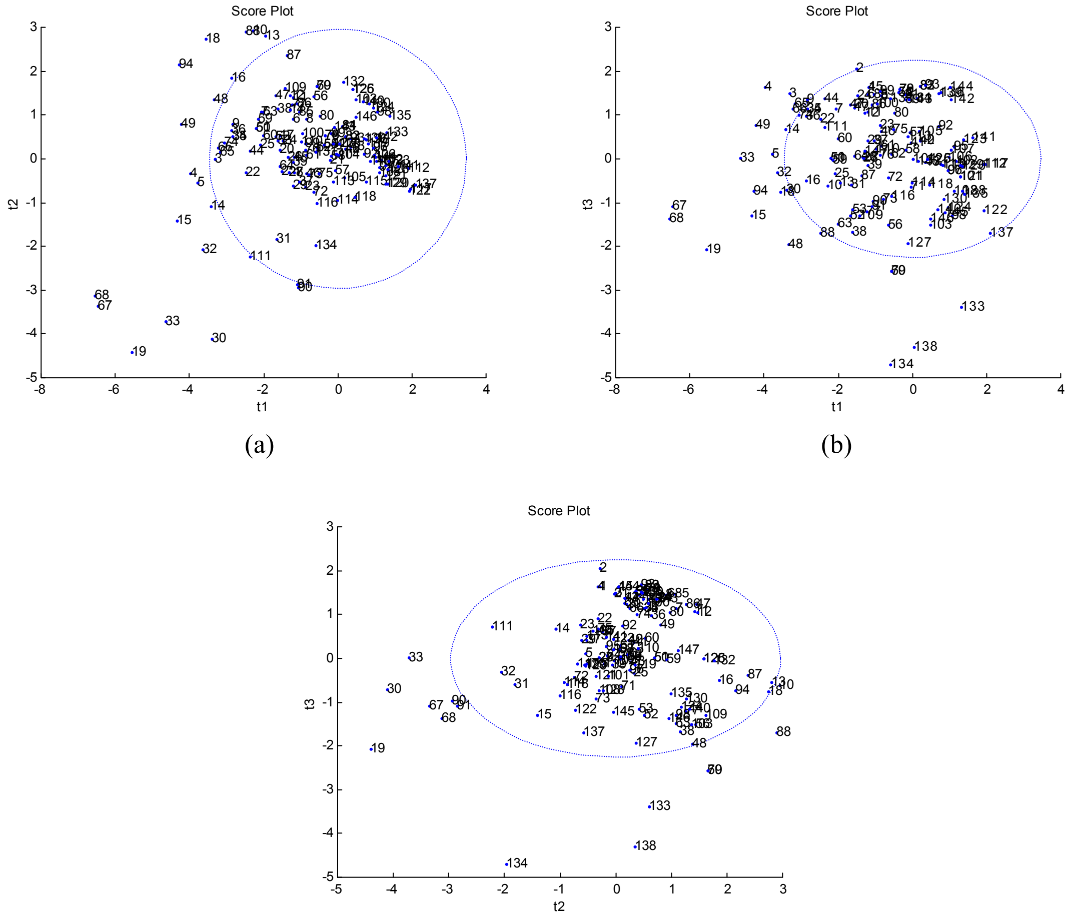
Normality plots of the in-control errors are displayed in Figure 5.7 and it can be said that errors are normally distributed except small deviations.



(a) (b) (c) (d) (e)

Figure 5.7. Normality Plots of In-control Errors in PCA Analysis of **Y** Data. Figure 5.7-a to Figure 5.7-e represent the normality of errors of variables 1 to 5, respectively.

Score plots (**t1**-**t2**, **t1**-**t3** and **t2**-**t3**) of **Y** matrix are illustrated in Figure 5.8. Plot of **t1** vs **t2** (Figure 5.8-a) states that observations 4, 5, 10, 13, 14, 15, 16, 18, 19, 30, 32, 33, 48, 49, 67, 68, 88 and 94, which are outside of the 95% control ellipse, are the possible out of control points. According to **t1**-**t3** and **t2**-**t3** plots (Figure 5.8-b and c), observations 3, 4, 5, 10, 15, 18, 19, 30, 32, 33, 48, 49, 65, 66, 67, 68, 69, 70, 88, 90, 91, 94, 133, 134 and 138 are expected as out of control.



(a)

(b)

(c)

Figure 5.8. Score Plots in PCA Analysis of **Y** Data. Figure 5.8-a to Figure 5.8-c represent **t1** vs **t2**, **t1** vs **t3** and **t2** vs **t3**, respectively.

T^2 plot of scores (Figure 5.9) gives some out of control points and they are the same as out of control alarms provided by score plots.

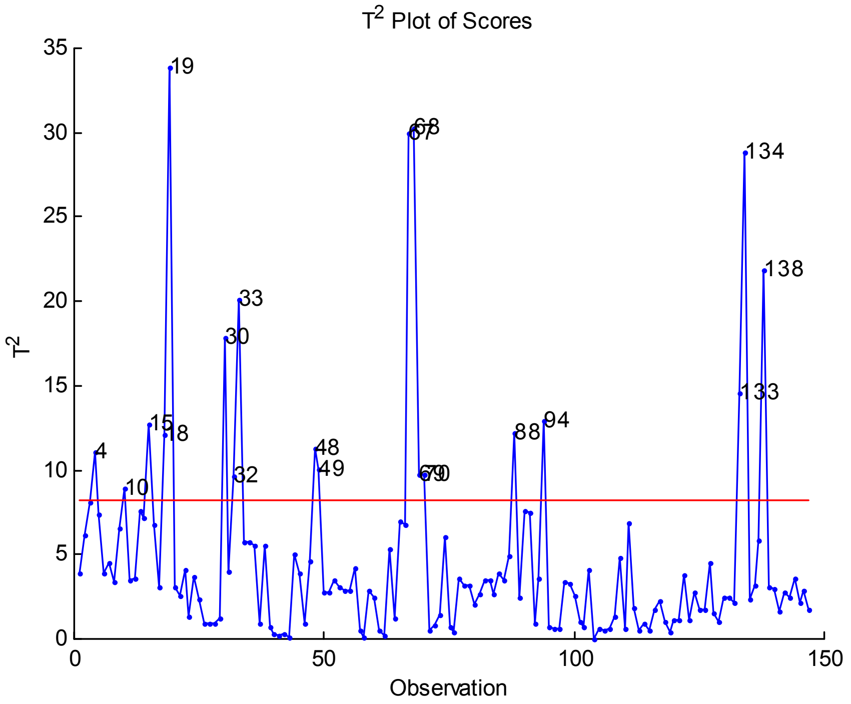


Figure 5.9. T^2 Plot of Scores in PCA Analysis of **Y** Data

Observations 19, 67, 68 and 134 are the most questionable out of control points. Their contribution plots are illustrated in Figure 5.10. The first contribution plot determines mold count (y5) and yeast count (y4) as the contributing factors to the out of control situation (Figure 5.10-a). The mold and yeast counts were recorded at observation 19. Contributing variables in observation 67 and 68 are the same, yeast count (y4), mold count (y5) and total viable count (y2) (Figure 5.10-b and c). Yeast counts of these observations are extremely high. Mold count and TVC also have high values. According to the last contribution plot (Figure 5.10-d), *E.coli* (y3), foreign material (y1) and mold count (y5) are responsible for the out of control state. *E.coli* and mold counts in observation 134 are high.

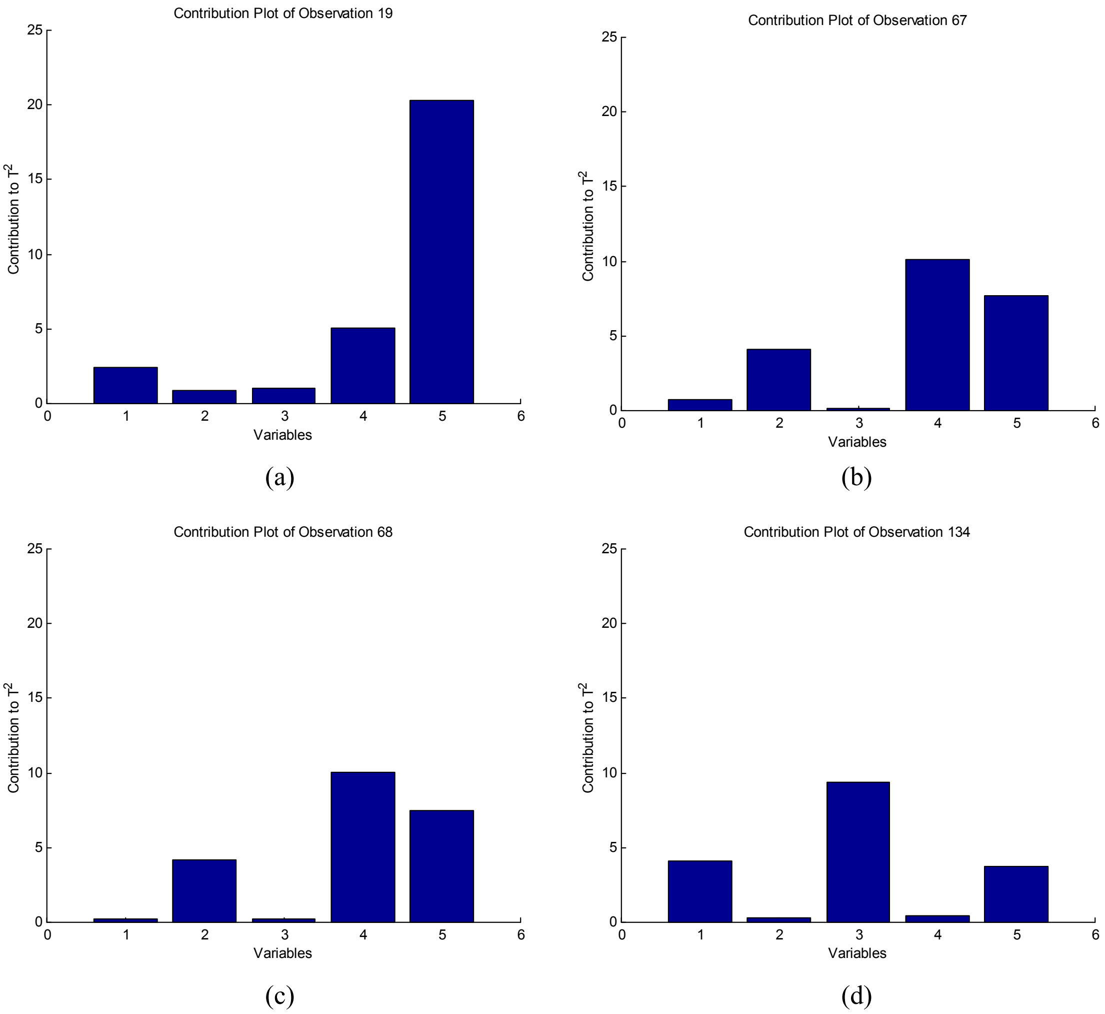


Figure 5.10. Examples of T^2 Contribution Plots in PCA Analysis of **Y** Data. Figure 5.10-a to Figure 5.10-d represent the contribution plots of observations 19, 67, 68, and 134, respectively.

Contribution plots of all out of control points in T^2 plot were analyzed in order to determine percentages of variable contributions (Figure 5.11). All of the variables have nearly the same percentage except y1. y1 does not contribute to the out of control points of T^2. y2, y3, y4 and y5 constitute microbial counts. Thus, microbial counts, which are the verification of CCP-1, are the most important contributing factor.

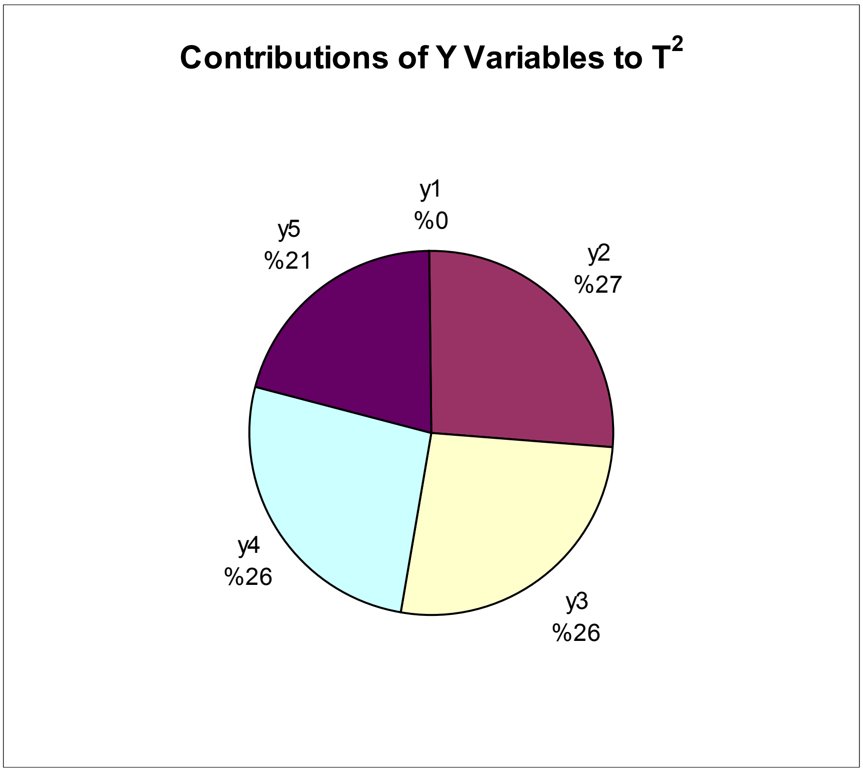


Figure 5.11. Contributions of **Y** Variables to T^2

The same singularity problem explained in the SPE computations of **X** matrix also appeared in that of **Y** matrix. Thus, the number of principal components increased to 5 in SPE computations of errors of **Y**.

SPE plot of **Y** is given in Figure 5.12. Out of control points of the SPE plot which was obtained by the PCA analysis of Y data are similar to the out of controls of T^2 plot. Some observations such as 7 and 38 are determined as out of control in SPE plot but not in T^2 plot, because these observations deviates from the model not from the mean value.

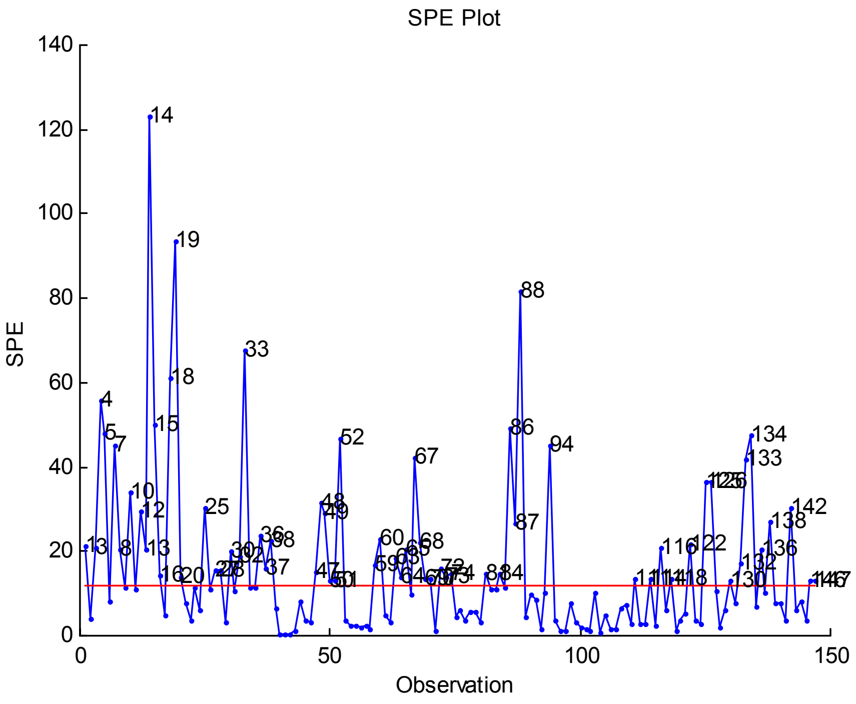


Figure 5.12. SPE Plot in PCA Analysis of **Y** Data

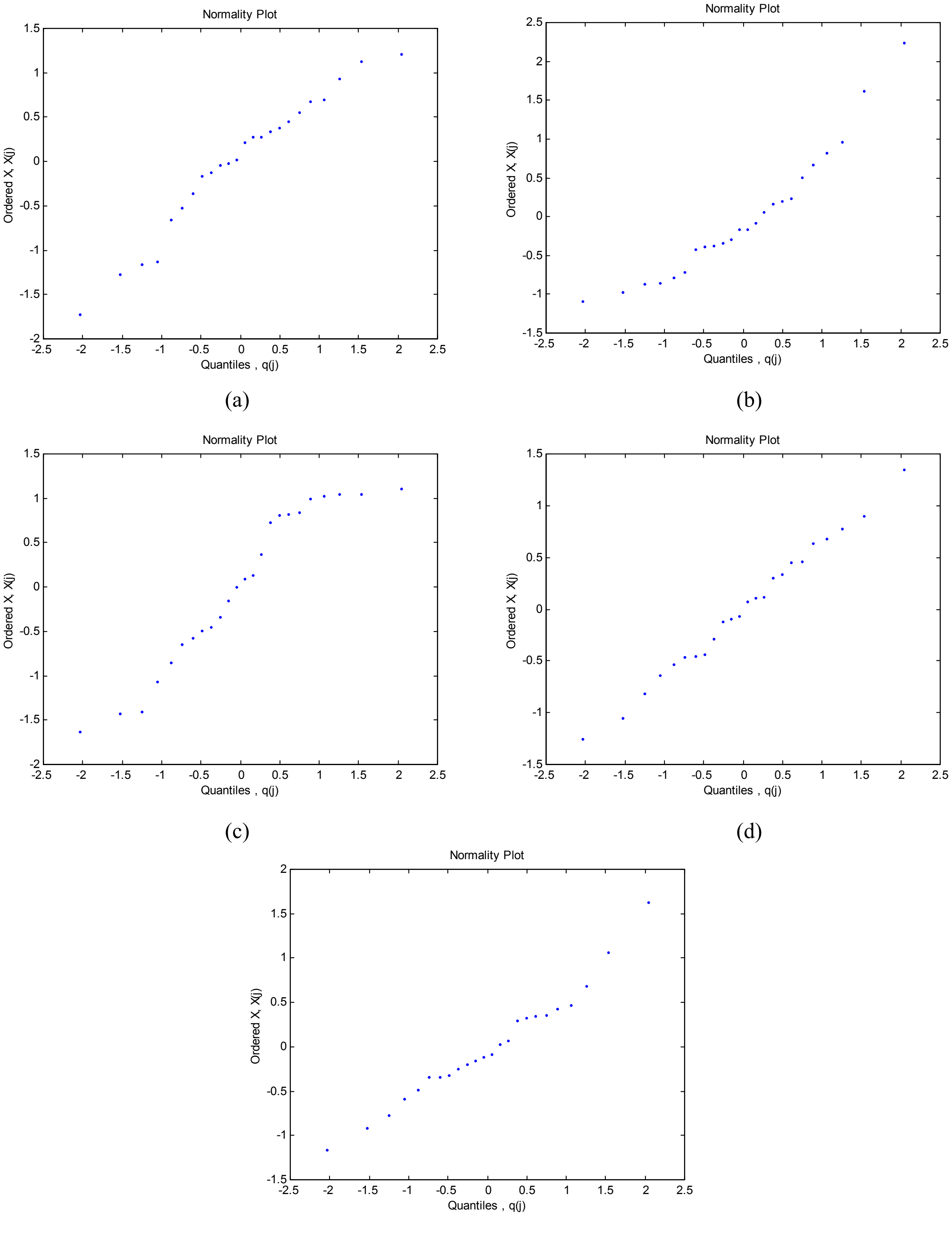
X matrix including raw material properties and process conditions and **Y** matrix including end product properties were analyzed by using different regression methods. The goal of using regression techniques was to investigate the relationship between **X** and **Y** matrices and to predict **Y** values by using the information obtained from in-control data set. An in-control data set was used to build the regression models and the model was applied to the entire data. **X** matrix was used directly without any transformation and 0.25 power transformation was applied to **Y** matrix in all regression models. The variables employed in the regression analyses are shown in Table 5.4.

Table 5.4. Variables of **X** and **Y** Matrices

Raw Material Properties	Plant origin foreign material	x1
	Rotten	x2
	Burst	x3
	Diseased	x4
	Crumpled	x5
	Brix	x6
Process Conditions	Blanching time	x7
	Blanching temperature	x8
End Product Properties	Foreign material	y1
	Microbial counts	
	Total Viable Count (TVC)	y2
	E. coli	y3
	Yeast	y4
	Mold	y5

5.3. Multiple Linear Regression (MLR)

In the MLR analysis, a regression model was built between **X** and **Y** variables. Normality of in-control errors and dependence of error values to **Y** were investigated. Normality plots of the in-control errors of **Y** are shown in Figure 5.13. Distribution of the variables are near to normality except the third **Y** variable (y3) as seen in Figure 5.13-c.



(a)

(b)

(c)

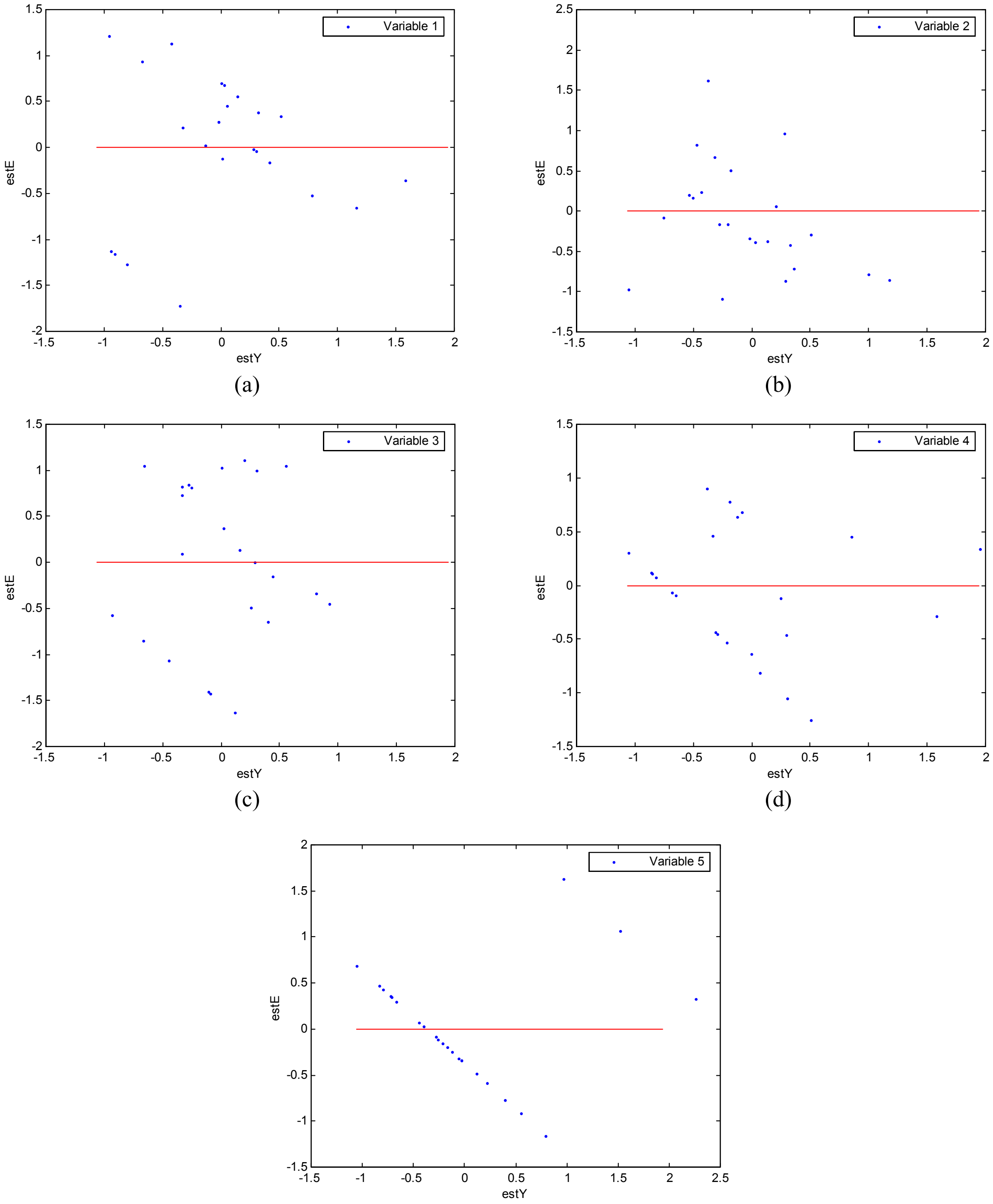
(d)

(e)

Figure 5.13. Normality Plots of In-control Errors of **Y** in MLR Analysis. Figure 5.13-a to Figure 5.13-e represent the normality of errors of variables 1 to 5, respectively.

Figures 5.14-a to 5.14-e represent the plots of estimated values of in-control errors versus estimated values of in-control **Y**. This plot is expected to be randomly distributed around zero mean if the model accurately explains the variability in the process. In

Figure 5.14-b, estimated y2 values (total viable count) are scattered enough versus its error counterpart (the second error component). However, Figures 5.14-a, c, d and e display that error values are still dependent to **Y** values and error values still have some information about the process. Thus, the model does not completely explain the relation between **X** and **Y**.



(a)

(b)

(c)

(d)

(e)

Figure 5.14. Plots of Estimated In-control Errors versus Estimated In-control **Y** Values in MLR Analysis. Figure 5.14-a to Figure 5.14-e represent the plot of estimated **Y** and **E** values of variables 1 to 5, respectively.

5.4. Principal Component Regression (PCR)

PCR model has 4 principal components explaining 80.99% of the variance of **X** data as in the case of PCA analysis. Therefore, error and score plots, T^2 and SPE graphs are the same as those in PCA analysis. Normality plots of in-control errors of **X**, score plots, T^2 plot of score matrix, contribution plots of T^2 out of control points and SPE plot of **X** are given in Figures 5.1, 5.2, 5.3, 5.4 and 5.6, respectively. T^2 and SPE charts revealed the expected out of control observations as stated in section 5.2 (Figure 5.15 and Figure 5.16).

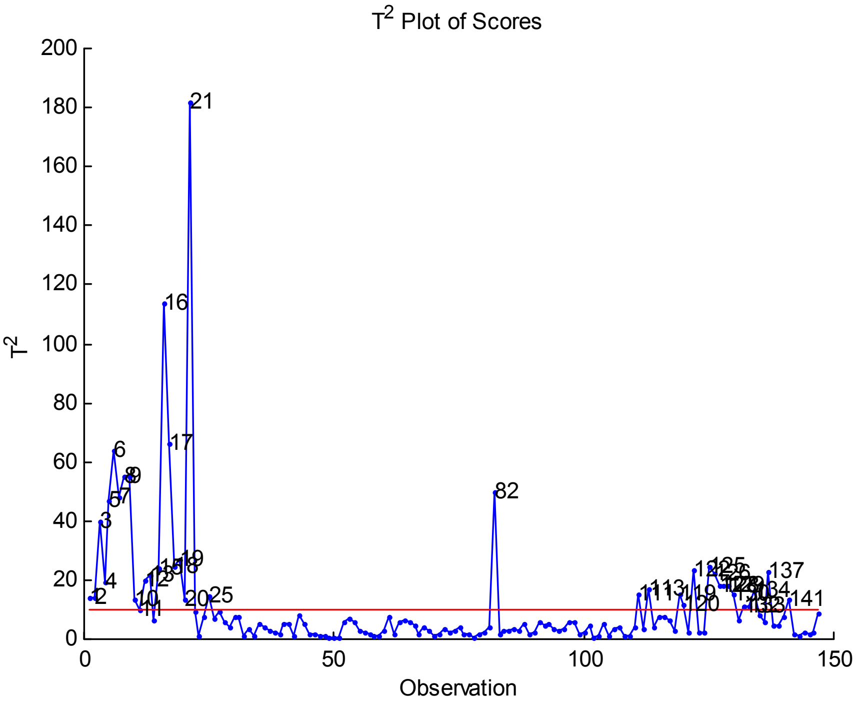


Figure 5.15. T^2 Plot of Scores in PCR Analysis

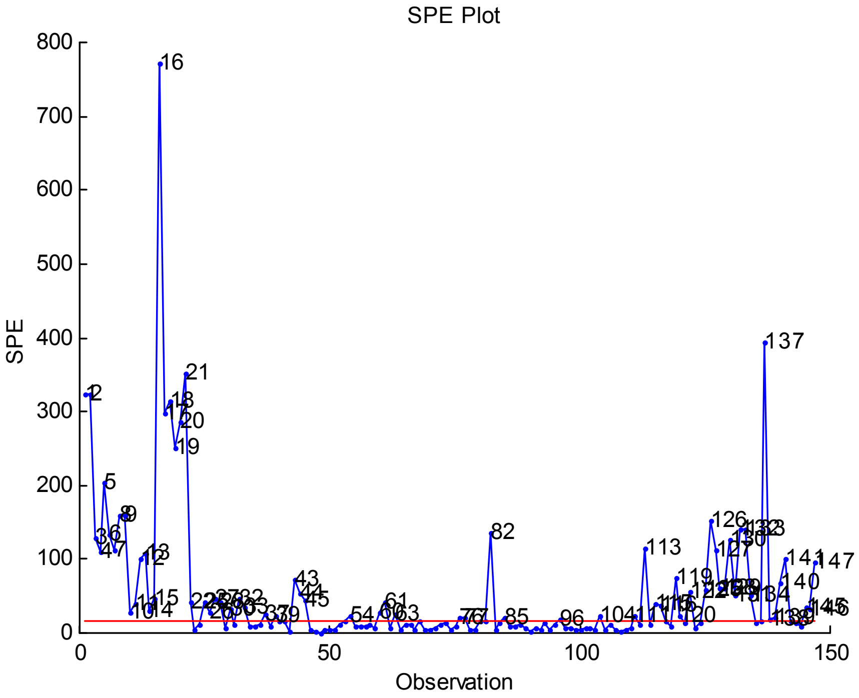


Figure 5.16. SPE Plot of Errors of **X** in PCR Analysis

Besides the results of **X** data matrix, the plots of **Y** matrix are displayed as well. Figure 5.17 gives the normality plots of 5 in-control error components. The model did not produce errors with normal distribution. In addition to this result, the plots of model errors versus predicted **Y** values show a dependency between these two estimations as in MLR model of **X** and **Y** (Figure 5.18).

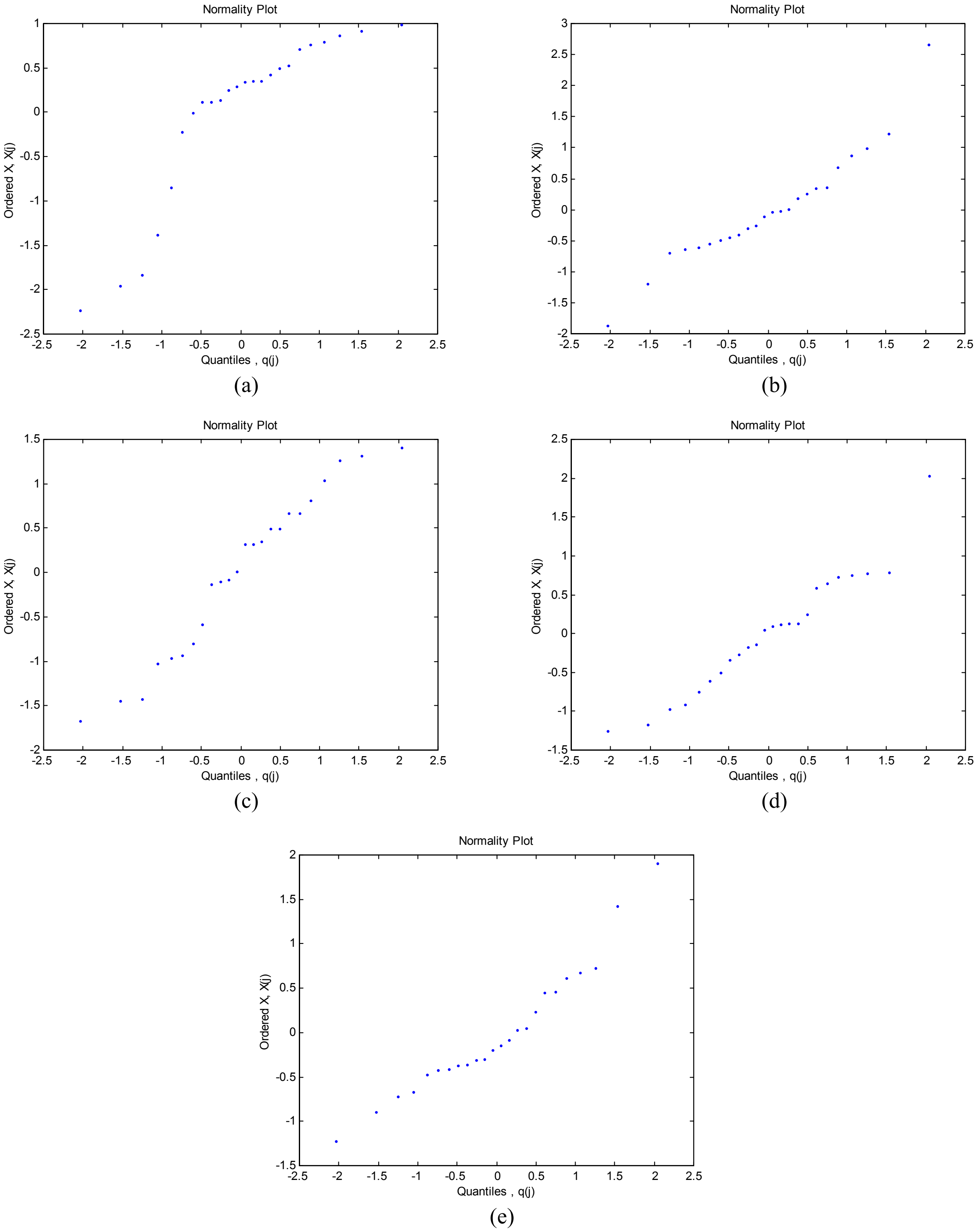
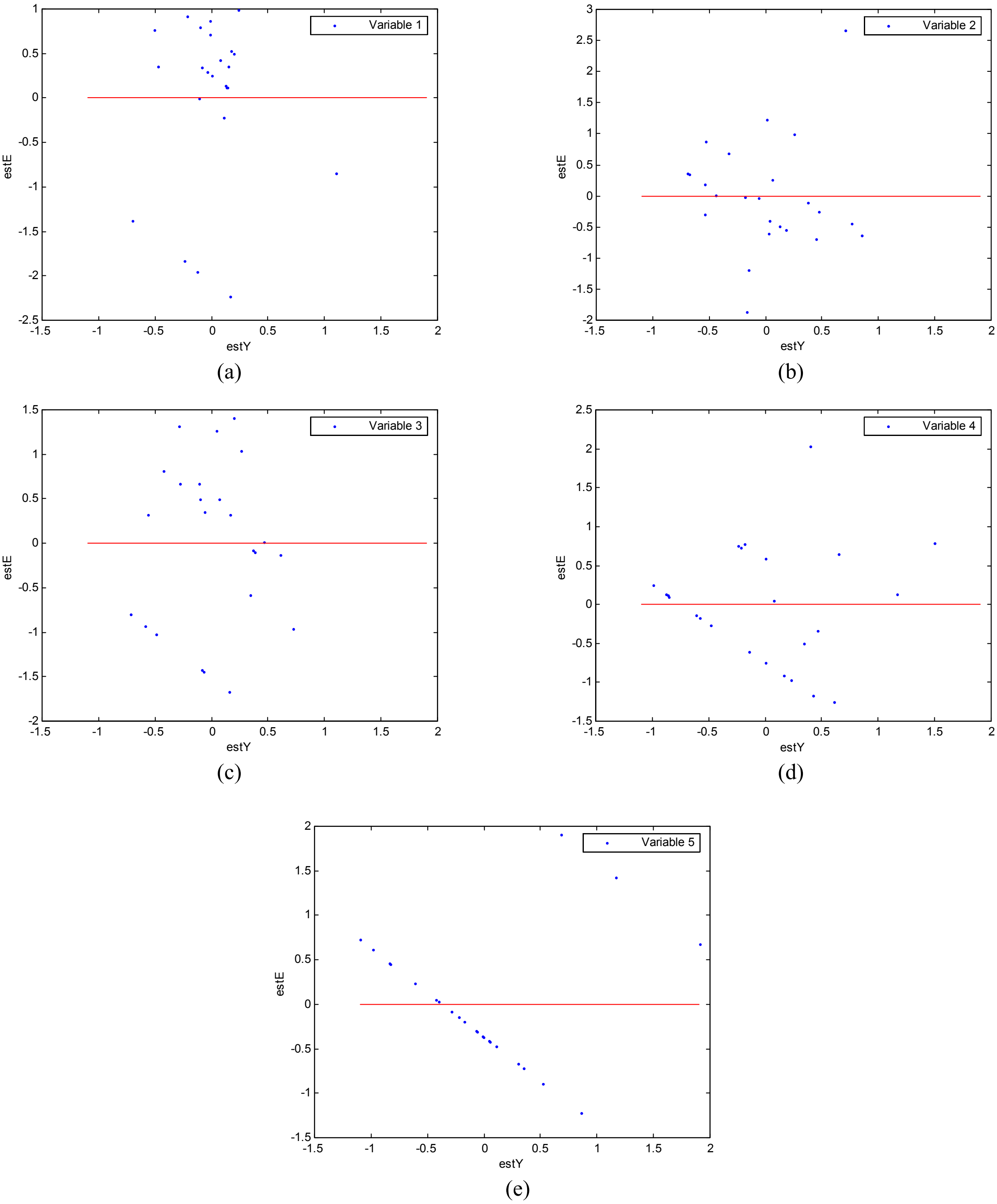


Figure 5.17. Normality Plots of In-control Errors of **Y** in PCR Analysis. Figure 5.17-a to Figure 5.17-e represent the normality of errors of variables 1 to 5, respectively.



(a) (b)

(c) (d)

(e)

Figure 5.18. Plots of Estimated In-control Errors versus Estimated In-control **Y** Values in PCR Analysis. Figure 5.18-a to Figure 5.18-e represent the plot of estimated **Y** and **E** values of variables 1 to 5, respectively.

SPE plot of **Y** is in Figure 5.19.

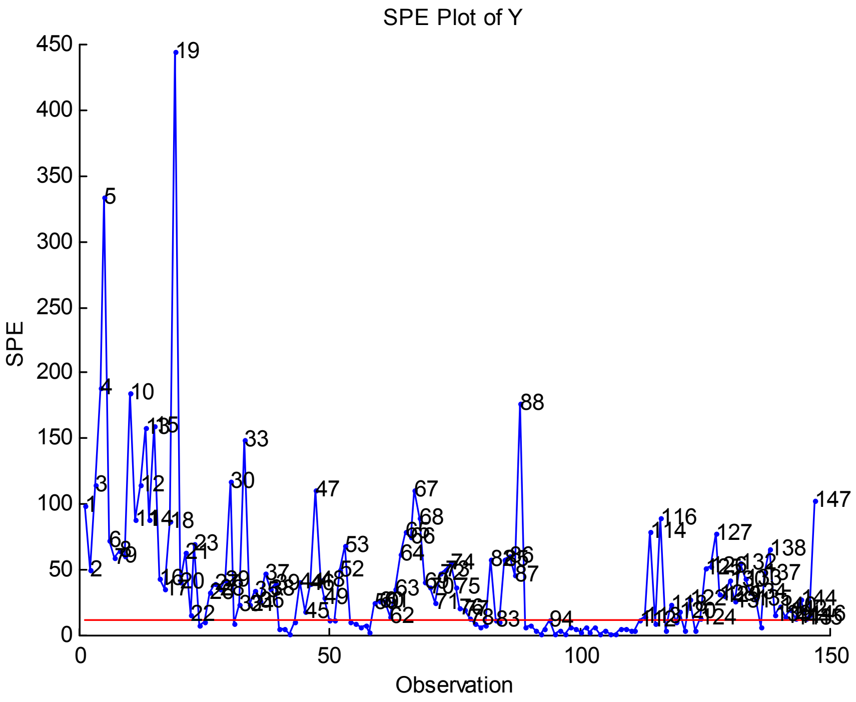


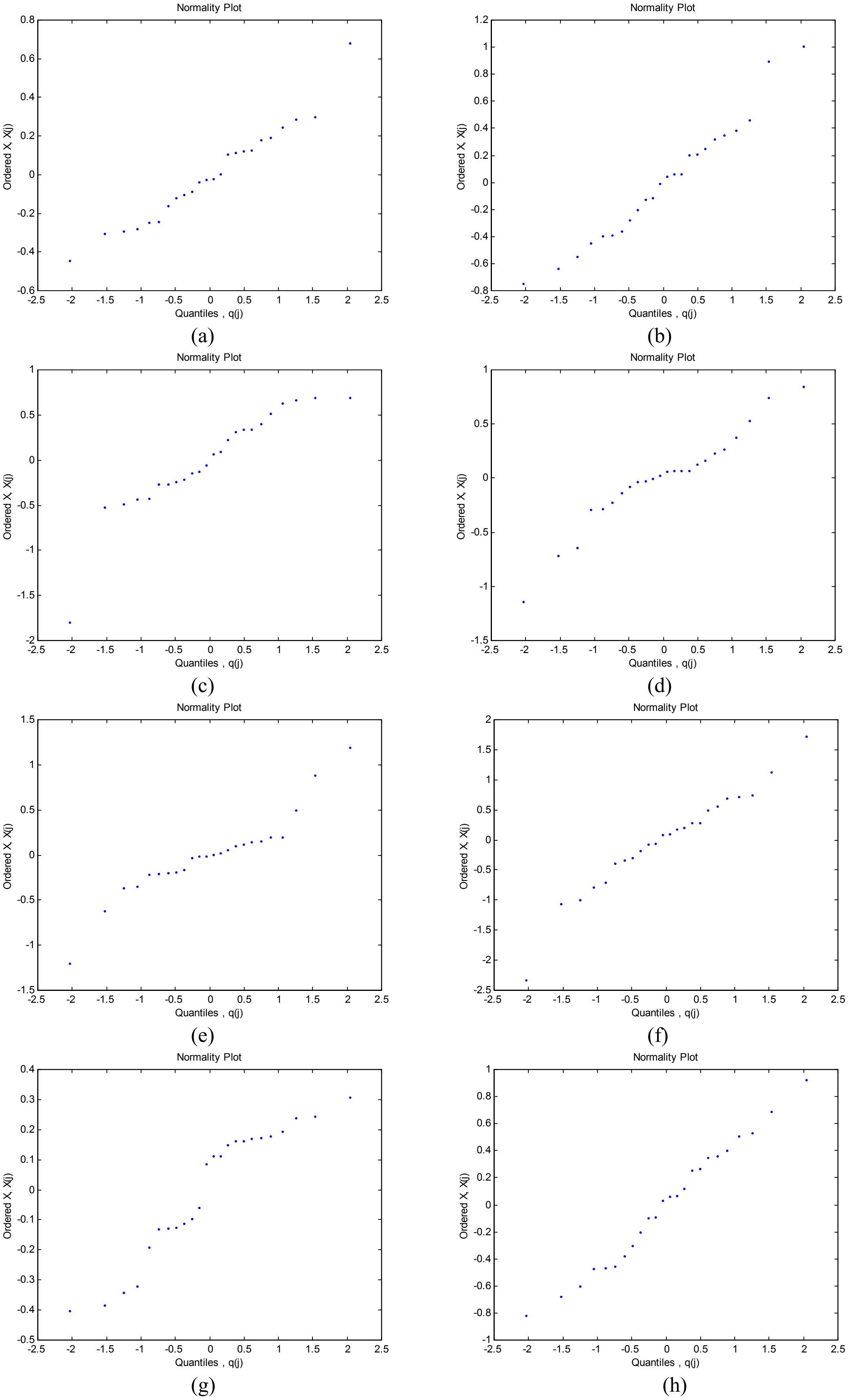
Figure 5.19. SPE Plot of Errors of **Y** in PCR Analysis

SPE plot of PCR analysis (Figure 5.19) displays a strong similarity to SPE plot in PCA analysis of **Y** data (Figure 5.12). In both figures, SPE plot gives out of control alarms where the differences between the actual and predicted values (errors) are high. Therefore, observations of **Y** data, which are higher than expected, are captured as out of control points. SPE results indicate that PCR method achieved good results by giving alarms for elevated values in the data set. PCR model could not extract the process information from the data and errors are still dependent to **Y** values as seen in Figure 5.18. Thus, PCR regression cannot produce acceptable result for prediction of **Y** data.

5.5. Partial Least Square Regression (PLSR)

In PLSR analysis, 4 principal components explaining 76.12% of the variance of **X** data and 34.82% of the variance of **Y** data were selected.

Normality plots of in-control errors of **X** and **Y** matrices are represented in Figure 5.20 and Figure 5.21, respectively. These plots also display some deviations from normality confirming that the PLSR model does not explain the variability in the **Y** data.



(a)

(b)

(c)

(d)

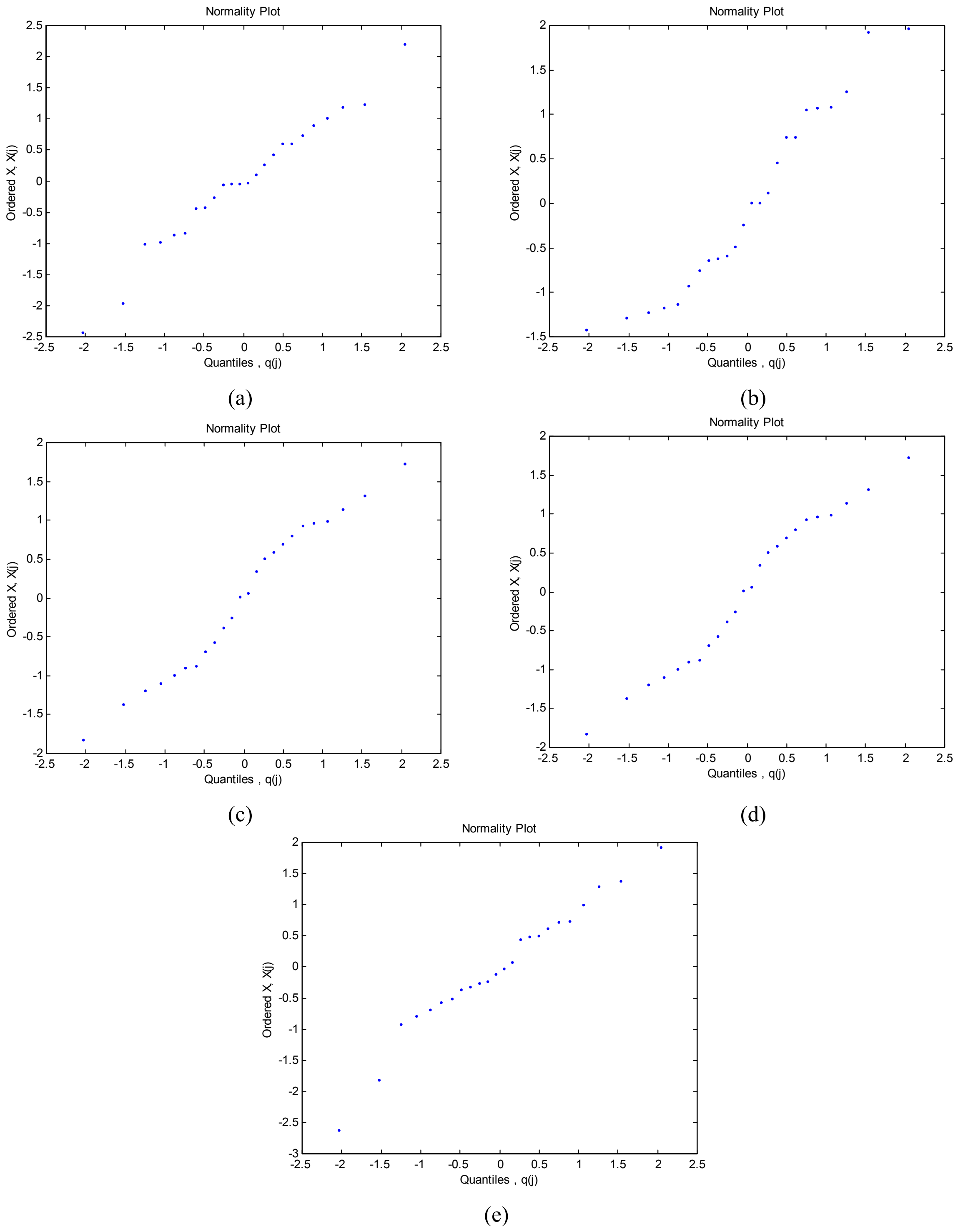
(e)

(f)

(g)

(h)

Figure 5.20. Normality Plots of In-control Error Components of **X** Data in PLSR Analysis. Figure 5.20-a to Figure 5.20-h represent the normality of errors of variables 1 to 8, respectively.



(a)

(b)

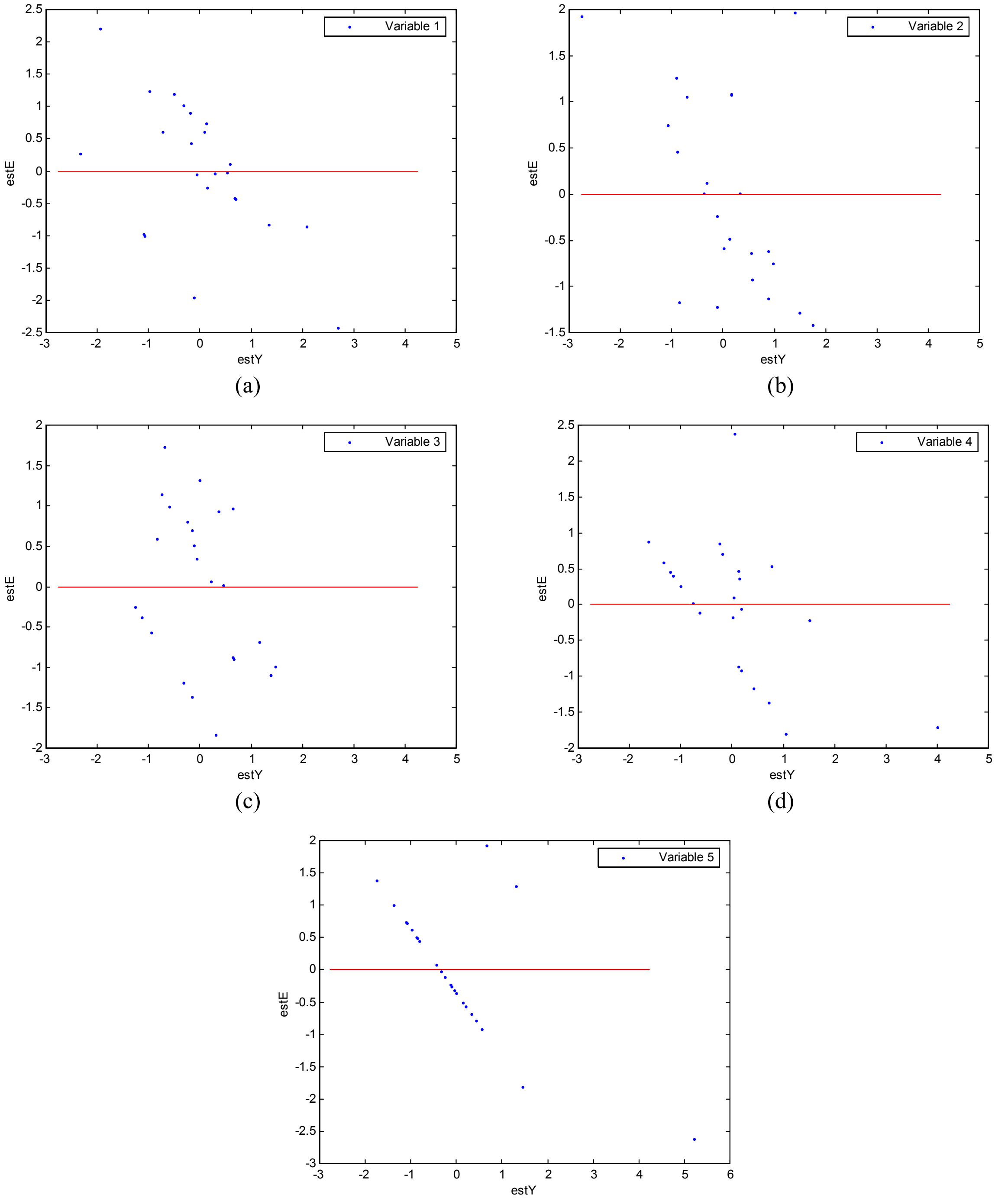
(c)

(d)

(e)

Figure 5.21. Normality Plots of In-control Error Components of **Y** Data in PLSR Analysis. Figure 5.21-a to Figure 5.21-e represent the normality of errors of variables 1 to 5, respectively.

Plots of estimated values of in-control errors versus estimated values of in-control **Y** values are displayed in Figures 5.22-a to Figure 5.22-e. In plots a, b and c, the data are scattered well. However, in plots d and e, the estimated values of variables y4 and y5 show a pattern which shows the insufficiency of the model in predictor.



(a)

(b)

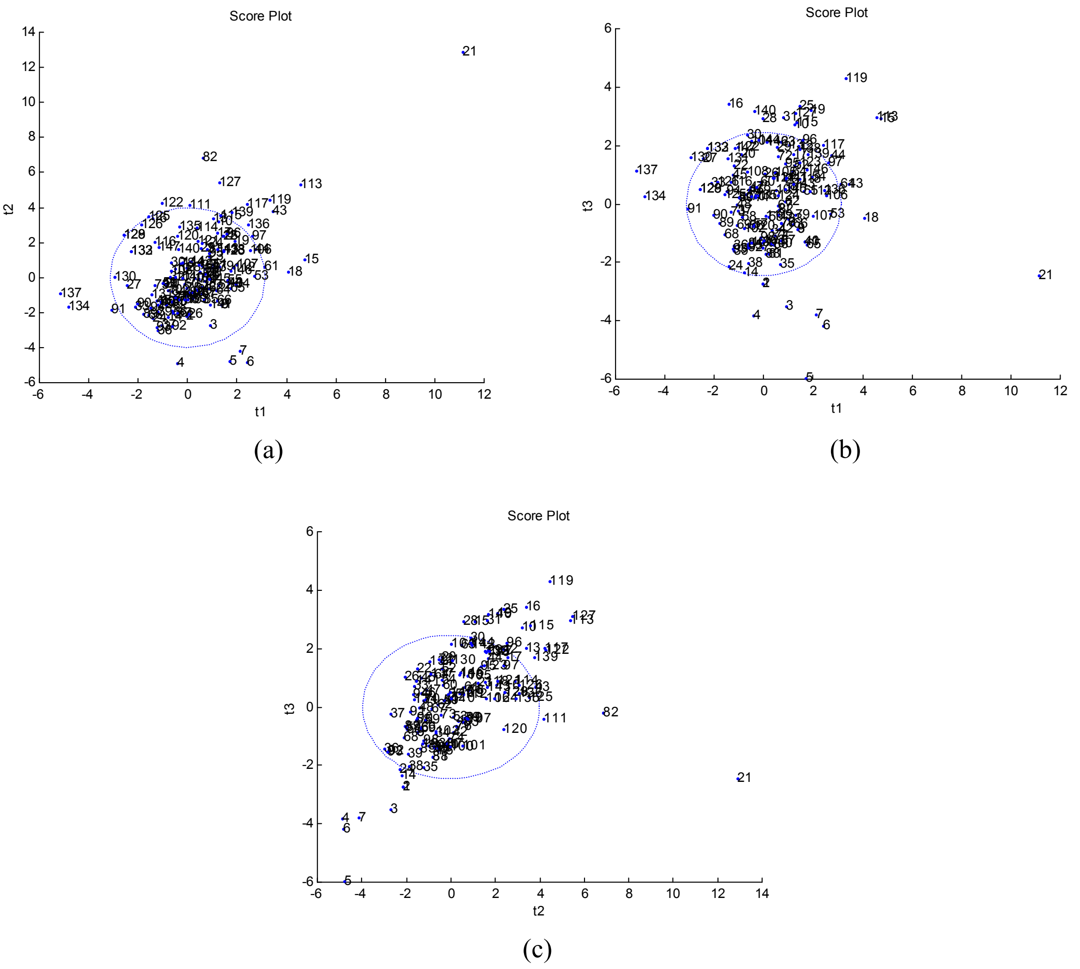
(c)

(d)

(e)

Figure 5.22. Plots of Estimated In-control Errors versus Estimated In-control **Y** Values in PLSR Analysis. Figure 5.22-a to Figure 5.22-e represent the plot of estimated **Y** and **E** values of variables 1 to 5, respectively.

Scatter plots of score components **X** and **Y** data are given in Figure 5.23 and Figure 5.24, respectively. In the score plots of PLSR model of **X** data (Figure 5.23), observations 1, 2, 3, 4, 5, 6, 7, 10, 13, 14, 15, 16, 18, 19, 21, 25, 28, 31, 43, 44, 82, 91, 96, 111, 113, 115, 117, 119, 122, 127, 130, 134, 136, 137, 139 and 140 show up as out of control points in this particular production season. T^2 plot of scores of the same model (Figure 5.25) proves that these observations also appear as the out of control signals. According to the score plots of **Y** data (Figure 5.24), points 5, 10, 13, 15, 18, 19, 30, 32, 33, 52, 53, 67, 68, 69, 70, 71, 88, 90, 91, 94, 111, 116, 118, 125, 126, 127 132, 133, 134, 138, 142, 144, 145 and 147 are at a distance from the confidence ellipse of in-control data points.

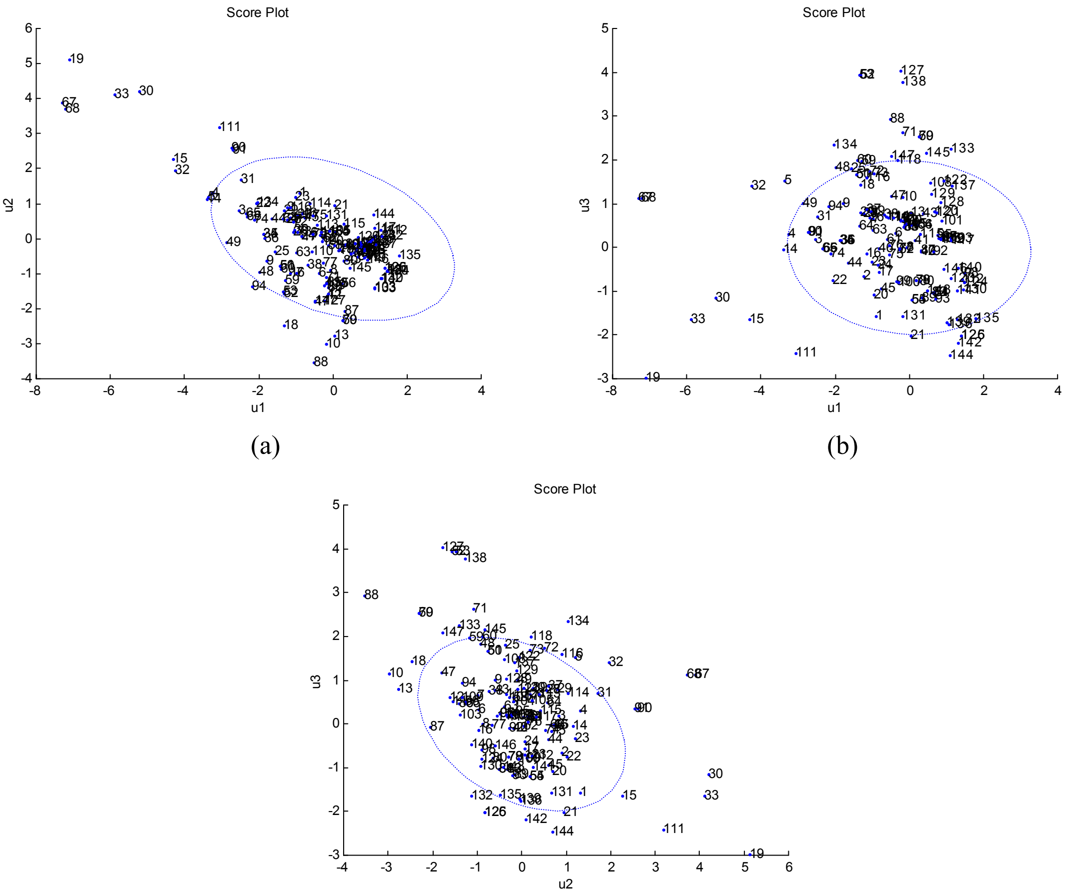


(a)

(b)

(c)

Figure 5.23. Score Plots of **X** in PLSR Analysis. Figure 5.23-a to Figure 5.23-c represent **t1** vs **t2**, **t1** vs **t3** and **t2** vs **t3**, respectively.



(a) (b)

(c)

Figure 5.24. Score Plots of **Y** in PLSR Analysis. Figure 5.24-a to Figure 5.24-c represent **u1** vs **u2**, **u1** vs **u3** and **u2** vs **u3**, respectively.

Independency between the score variables of PLSR analysis of **X** and **Y** data could not be assured as seen in Figure 5.23 and Figure 5.24. Especially, the scatter plot of **t2** versus **t3** (Figure 5.23-c) and **u1** versus **u2** (Figure 5.24-a) shows the trend between two score variables. This explains the insufficiency of the model to reveal the information between and within **X** and **Y** data sets.

T^2 plots of **X** and **Y** scores (**T** and **U**) are represented in Figure 5.25 and 5.26. The same picture obtained by the score plots is observed in the T^2 plot of **Y** scores (Figure 5.26).

T^2 plot of **X** scores in PLSR analysis (Figure 5.25) is similar to T^2 plot of scores in PCA analysis of **X** data (Figure 5.3). In the figures, out of control points in **X** scores are captured successfully. These out of control points are the observations which **X** variables collectively produce high values in the scores. As in the results of **X** data, T^2 plot of **Y** scores in PLSR analysis (Figure 5.26) shows the same trend with T^2 plot of scores in PCA analysis of **Y** data (Figure 5.9). Observations which have variables producing high score values simultaneously are captured as out of control alarms. Thus it is possible to say that, T^2 plots of PLSR method successfully analyzed **X** and **Y** data individually. However, the PLSR technique did not guarantee a good regression model for **X** and **Y** data sets.

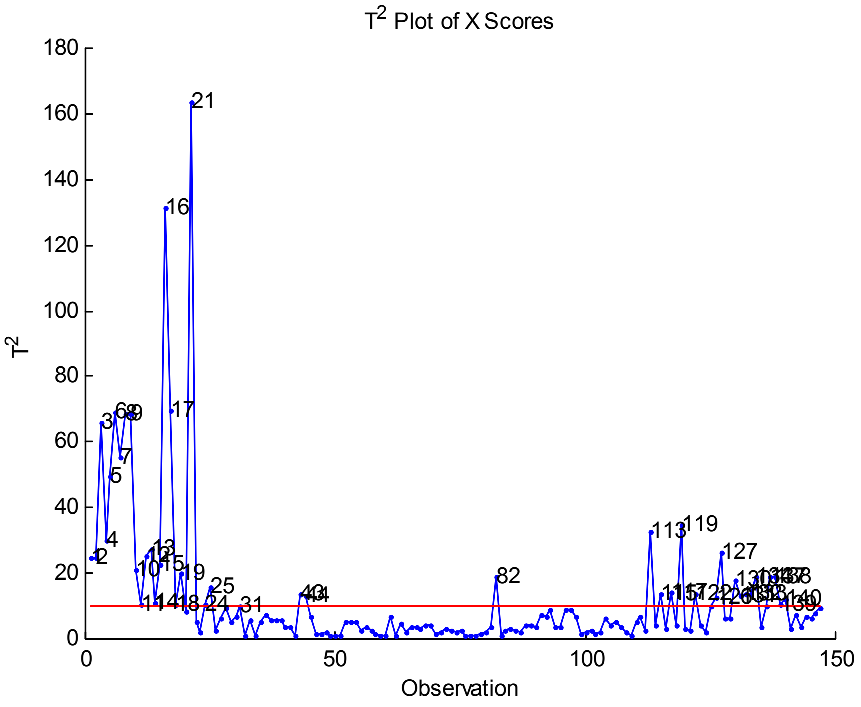


Figure 5.25. T^2 Plot of **X** Scores in PLSR Analysis

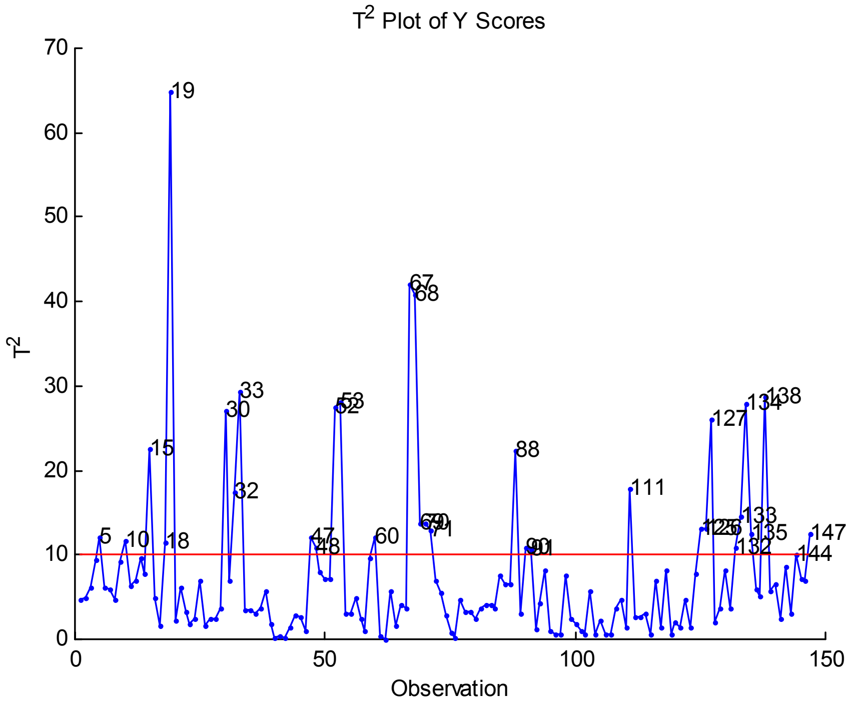


Figure 5.26. T^2 Plot of **Y** Scores in PLSR Analysis

Figure 5.27 illustrates the SPE plot of errors of **Y** in PLSR analysis. General trend and the out of control alarms of this plot are quite similar to SPE plot in PCA analysis of **Y** data (Figure 5.12) and SPE plot of PCR analysis (Figure 5.17).

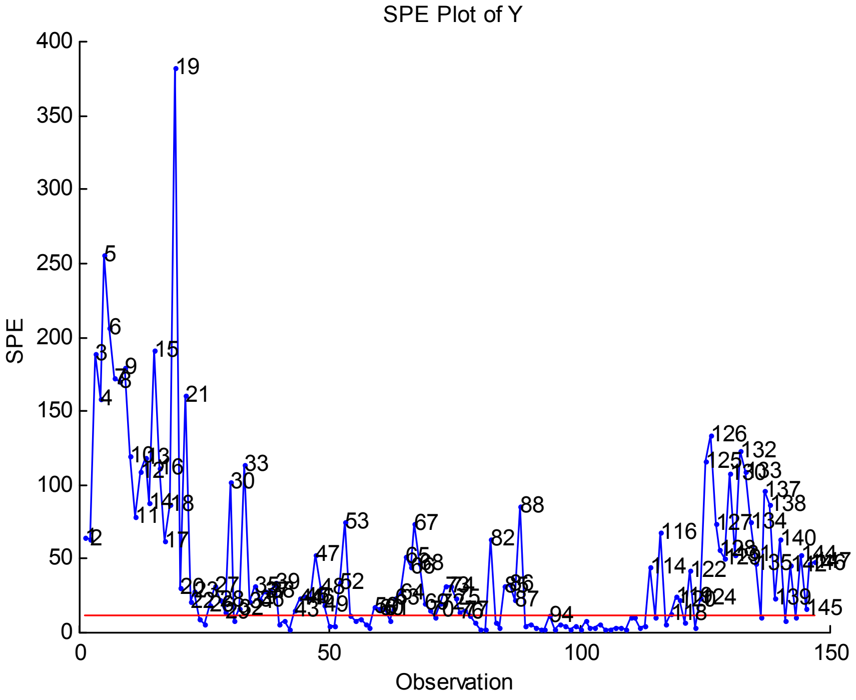


Figure 5.27. SPE Plot of PLSR Analysis

5.6. Performance of Multivariate SPM Techniques in Analyzing the Process

In the PCA analysis of **X** data, the trend of raw material and process variables during the process was successfully observed by T^2 and SPE plots. PCA model of **X** data determined the points, which are high in the real data, as out of control points in T^2 and SPE plots as expected. T^2 contribution plots were good at identifying the contributing variables. According to the monitoring plots, out of control points accumulates at the beginning and at the end of the production. This trend may appear because of the seasonal changes in the raw material or in the process. The properties of the vegetable may change during the season. At the beginning and at the end of the season, raw material incoming to the plant is not continuous. Thus, the storage time of raw material is long. However, in the midseason, raw material is continuously processed without storage. When the raw material input rate is continuous, only red pepper is processed in the process lines. At the beginning and at the end, the amount and type of the raw material, which is to be processed, change frequently. PCA model of **Y** data also explained the end product variables well. Statistical monitoring charts gave correct alarms for high values of observations. T^2 contributing plots were also successful in defining the contributing variables.

MLR method could not extract the model from the in-control data set as expected, since MLR technique may not give good results with correlated data as mentioned in Section 3.2.2.2. The inability of MLR to form the model to explain the correlation between **X** and **Y** was an expected result since the data of the study was correlated.

PCR model plots analyzed **X** and **Y** data individually well. T^2 plot of scores and SPE plot determined the expected points as out of control. However, the model did not produce errors independent of **Y** values. Thus, the relation between **X** and **Y** data could not be formed. It is known that PCR method is insufficient when compared to PLSR method since it produces uncorrelated scores only for **X** data but does not consider the information in **Y** data. Thus, PCR results are not unexpected.

In PLSR analysis, T^2 plots of **X** and **Y** data successfully determined the high values of the data set by out of control alarms. Thus, the analysis of the two data sets was quite satisfactory. However, the model explained 34.82% of the variance of **Y** data. This percentage was not enough to find out the relation of **X** and **Y** data. Also, errors still had the process knowledge and the model was not constructed well. PLSR is known as an efficient regression technique in the literature. It considers both **X** and **Y** scores and constructs a very strong model to explain the correlation of two data sets. In this study, PLSR technique did not produce satisfying results for modeling purpose. These sorts of problems may arise due to the insufficient information (dynamic) in the in-control data set.

In-control data is the data collected from a process under normal operating conditions. The in-control data or modeling data were selected from the data collected during the production season. Thus, the in-control data set was not the one, which explains the normal operating conditions exactly. Another drawback here is that some very important process variables could not be used and some of the process information lost.

5.7. Contribution of Multivariate SPM Methods and Tools to HACCP Program

In this study, only the data of CCP-1 and CCP-3 were analyzed statistically since the data of CCP-2, which are the temperature measurements of IQF, are not collected by the plant and the data of CCP-4, which are the metal detector results, are not the same type with the other variables. The data of CCP-1 were process conditions; blanching time and blanching temperature and the data of CCP-3 were foreign materials. **X** matrix included the data of CCP-1 (x7 and x8) and raw material properties (x1 to x6) which are important parameters to decide the process conditions. **Y** matrix included the data of CCP-3 (x9) and microbial counts (x10 to x13) which are the verification of CCP-1; blanching.

X matrix composed of process related data and **Y** matrix composed of product related data were analyzed individually by PCA method and analyzed in cause and effect relationship by regression methods. The main aim of using multivariate methods was to obtain uncorrelated scores of the models since it is not suitable to analyze the correlated data including many variables by univariate SPM methods. These scores were investigated by multivariate statistical tools; T^2 and SPE charts, in order to determine the out of control points of the process. Contribution plots were used to find the contributing variables to these out of control points. Contribution percentages of variables obtained from out of control points of T^2 were calculated.

PCA results displayed the most problematic variables of process and product data (**X** and **Y** matrices). The contribution percentage of the variables obtained from out of control points of T^2 plot of **X** data pointed out plant origin foreign material (x1) and blanching temperature (x8) as the most important contributing factors. The former indicates the effect of the raw material quality and the latter proves the role of CCP-1, which is blanching. The contribution percentage of the variables obtained from out of control points of T^2 plot of **Y** data showed that the microbial counts (y2, y3, y4 and y5) which are the verification of CCP-1 are completely responsible from the out of control cases. These results indicate that CCP-1 (blanching) is the point in which extra care should be taken. The increased attention and control of this CCP will result in a safer and more effective process. Other factor that needs attention and affects the microbial load of the end product is the acidity of washing water used almost in every step of production.

Three different regression techniques were used to relate raw material properties and process conditions (**X** data) to end product variables (**Y** data). The aim of using regression analysis is to forecast the end product properties beforehand and take the necessary corrective actions earlier without waiting for the result of long procedure of microbial analyses.

Since MLR regression technique does not remove the correlated structure in process data, it is not possible to use the multivariate monitoring charts. The quality variables (**Y** data) may be monitored univariately (separate monitoring charts for each variable). PCR and PLSR analysis, in which independent components are produced, are recommended in modeling of correlated process data. The new independent variables are used in multivariate process monitoring charts. In PCR analysis, the principal components (factors) are extracted from raw material and process variables (**X** data) independently of the quality variables (**Y** data), which are to be predicted. On the other hand, in PLSR analysis, the factors are computed with the information coming from both **X** and **Y** data sets. In other words, PLSR algorithm performs the principal component calculations to account the variation in raw material and process data while assuring that these new orthogonal variables relate to the end product variables (**Y**). This is to achieve better correlations to predicted variables (**Y**).

5.8. Cause and Effect Diagrams

Cause and effect diagrams which were performed for all CCPs individually are shown in Figures 5.27 - 5.30.

The first CCP is at blanching section. The risk of concern is insufficient thermal treatment. Cause and effect diagram for CCP-1 (Figure 5.27) displays the possible causes of the risk, which may arise from the system, procedures and equipment. System problems can be the increase of microbial load by keeping raw material waiting and contamination of product in the production line. The amount of the raw material transported to the plant changes with the season. When the amount of incoming raw material is too small or too high, it should be kept waiting since a certain amount of raw material should be loaded to the system each time. Therefore, the increase of microbial load is inevitable at the beginning and at the end of the production season. On the other hand, incoming material is continuous in the middle of the season and the problem under consideration does not occur. Microbial load may also increase because of the contamination of the product during the process. In these cases, blanching time and temperature would not be sufficient to decrease the microbial load to the desired level. Procedures such as washing and cutting may also cause a hazard. The vegetables are cut at certain dimensions after blanching and cutting errors are maximum at high blanching temperatures since the product is very tender. The temperature, which is low enough to avoid this problem but high enough to decrease microbial load, should be utilized. However, temperature level may not be optimized and insufficient heat treatment occurs. The same problem arises if the feed rate is high and the heat treatment may not be sufficient for the amount of product loaded. The product is washed with chlorinated and acidified water several times during the production and this is the first and only hurdle against microorganisms before blanching. If the amount of chlorine and acid in washing water is not enough, the microbial load would increase. The condition and settings of blanching equipment also influence the effectiveness of the heat treatment.

The second CCP is freezing of product since insufficient freezing is the risk of hazard. The possible causes are seen in Figure 5.28 and they are due to the procedures, equipment and system. Defrost of IQF is very important for the effectiveness of the freezing process. An error or delay of defrost timing causes insufficient freezing. Freezing is also effected by high feed rate as similar to blanching step. The amount of the product loaded to the equipment should be compatible with the freezing rate of the equipment. IQF settings should be arranged according to the product structure. Cutting size or pulp thickness may be the important parameters for the settings. Stoppage or freezing of IQF honeycombs decrease the efficiency of the equipment and freezing may be insufficient as a result of this. Similarly, efficiency of compressor is directly related with the efficiency of IQF since the incoming cooling material is controlled by the compressor.

Final grading and control is CCP-3 since presence of foreign materials in the product is the risk at this point. Figure 5.29 shows the possible causes of foreign material presence. There may be many factors resulting with this problem such as system, equipment, material, personnel, plant and surroundings. The major problem is insufficiency of the number of workers, which control the presence of foreign material at this step. Other problems are the foreign materials coming into the product from various origins. Foreign materials may be plastic piece from the equipment, stone, rope from the material, hair from the personnel, floor coating piece from the plant and bird hair, insect from the surroundings.

The last CCP is metal detector. The risk of concern is the presence of metals in the product. Cause and effect diagram of CCP-4 is displayed in Figure 5.30. Problems arising from the measurement such as breakdown of metal detector, decrease in accuracy and calibration problems are the major causes. Therefore, efficiency of the detector should be controlled regularly. Other causes of metal presence are equipment, personnel and surroundings. Metals which may be found in the product are screw, nail from the equipment, watch piece from the personnel and wire piece from the surroundings.

The use of cause and effect diagrams can be explained better by giving the example below. Let's consider a case in PCA analysis of process data. T^2 plot (Figure 5.3) detects observation 6 as out of control point and its contribution plot (Figure 5.4) determines blanching temperature (x8) as the main factor of this out of control point. Blanching operation is CCP-1 and its variables are blanching time and temperature. Blanching temperature is 85^oC at observation 6 and this is the minimum value of x8. The problem at this point may be insufficient thermal treatment and the possible causes of insufficient thermal treatment were determined by the cause and effect diagram of CCP-1 (Figure 5.27). Thus, cause and effect diagrams can be used as a problem analysis tool to identify the root causes of a particular problem.

5.9. Recommendations to On-site Data Collection in Frozen Food Production

PCA method was successfully performed for the statistical analysis of process data (**X**) and quality data (**Y**) individually. Regression methods which are MLR, PCR and PLSR determined the out of control points of the process, however they could not fully explain the correlation between **X** and **Y**. It was concluded that the reasons could be linked to the collection of data in the production line.

No major problem was observed when these data sets are analyzed individually since they are consistent by themselves. However, raw material and process data and quality data do not appear to be complementary when their relation is analyzed. For instance, the correlation between blanching and microbial counts can not be explained. This makes sense when it is realized that **X** and **Y** data do not match each other.

Raw material, process and end product data are documented in corresponding departments in the plant. The measurements at the control points are recorded on special data forms during the production. The recording procedure has to be done by personnel authorized for this duty at each shift. The same person should copy the data to a network computer in which all the information from other departments will be collected simultaneously.

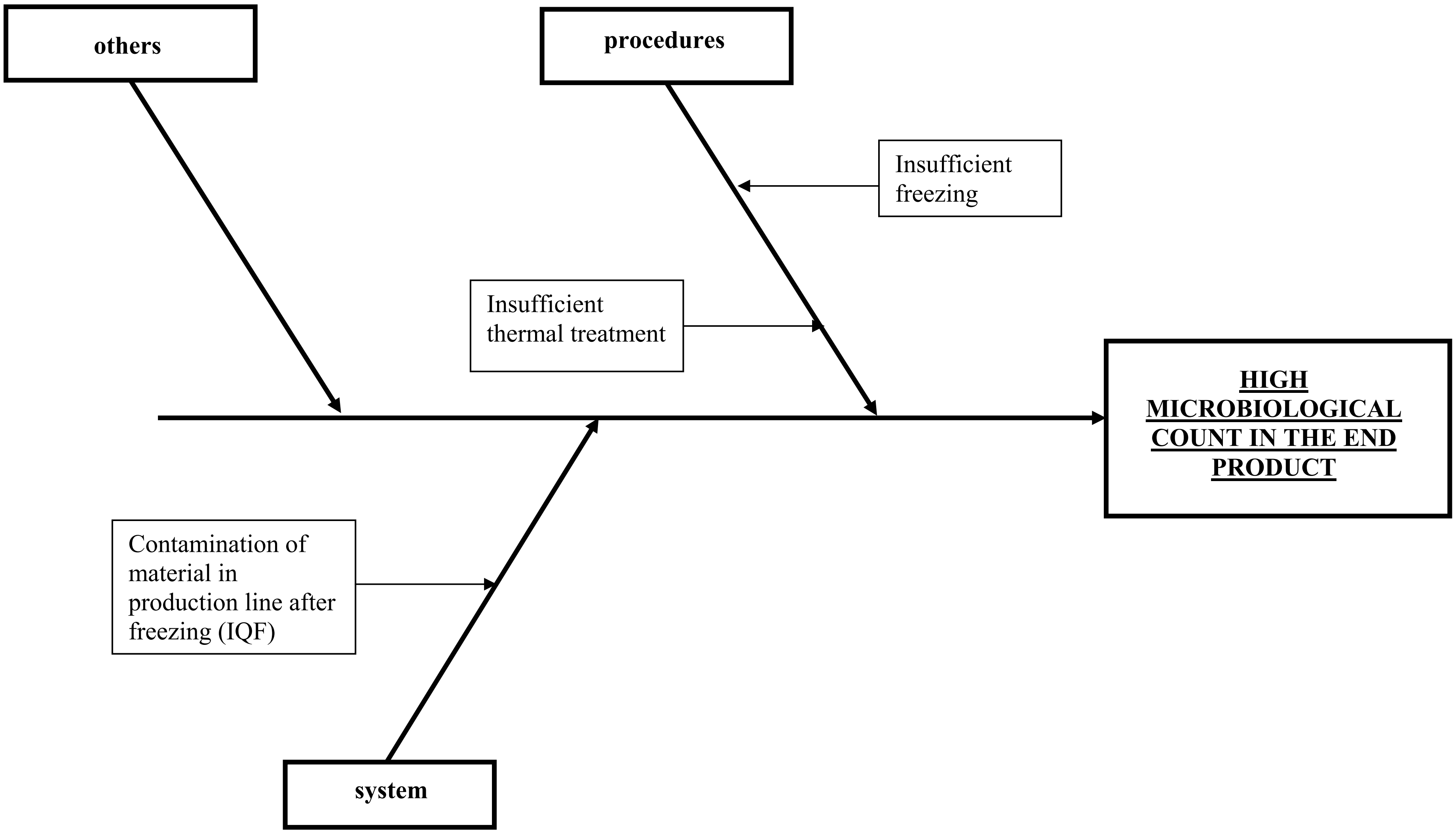
A few process variable, which would carry quite amount of information, are neither measured nor recorded at the same sampling rate as the other variables. Storage time of the vegetable or fruit is the time elapsed between the arrival of the raw material and its processing. It should be recorded on the data forms. pH of the washing water and the product temperature at the exit of IQF would be very valuable as process variables. Chlorine and citric acid concentrations in the washing water are collected to adjust chlorine level and acidity of washing water at certain times. If they were recorded simultaneously as other variables, the statistical models would be much more effective both in prediction and monitoring.

Most of the time full automation is not possible at every production stage such as sorting the raw material on conveyor belts to remove foreign materials, analyzing incoming raw material for diseased and rotten parts or microbial analysis of the end product. In this case, human factor is very important in collecting, analyzing and documenting data. Maximum attention should be paid.

Some recommendations can be made with the experience of all these problems for a study with the aim of statistical analysis or modeling of data, which belong to a real process ;

- All collected variables, which have the information on the process, should be complete, have the same period, complementary of each other and traceable.
- Human factor on collecting data should be reduced to minimum by providing automation.
- The personnel who is responsible from the documentation should be educated. Data forms should always be recorded by an authorized personnel and not change hands.
- Data forms should be clear, easy to understand and easy to pursuit.
- Different departments should collect and document their data complementary of each other for traceability.

A more effective statistical analysis and process modeling can be performed with the data collected as recommended above. The knowledge on the process obtained from this study can be used as feedback in the following production seasons. This will enable a safer and more effectively controlled production.



others
procedures
Insufficient freezing
Insufficient thermal treatment
HIGH MICROBIOLOGICAL COUNT IN THE END PRODUCT
Contamination of material in production line after freezing (IQF)
system

Figure 5.28. Cause and Effect Diagram for High Microbiological Count in the End Product

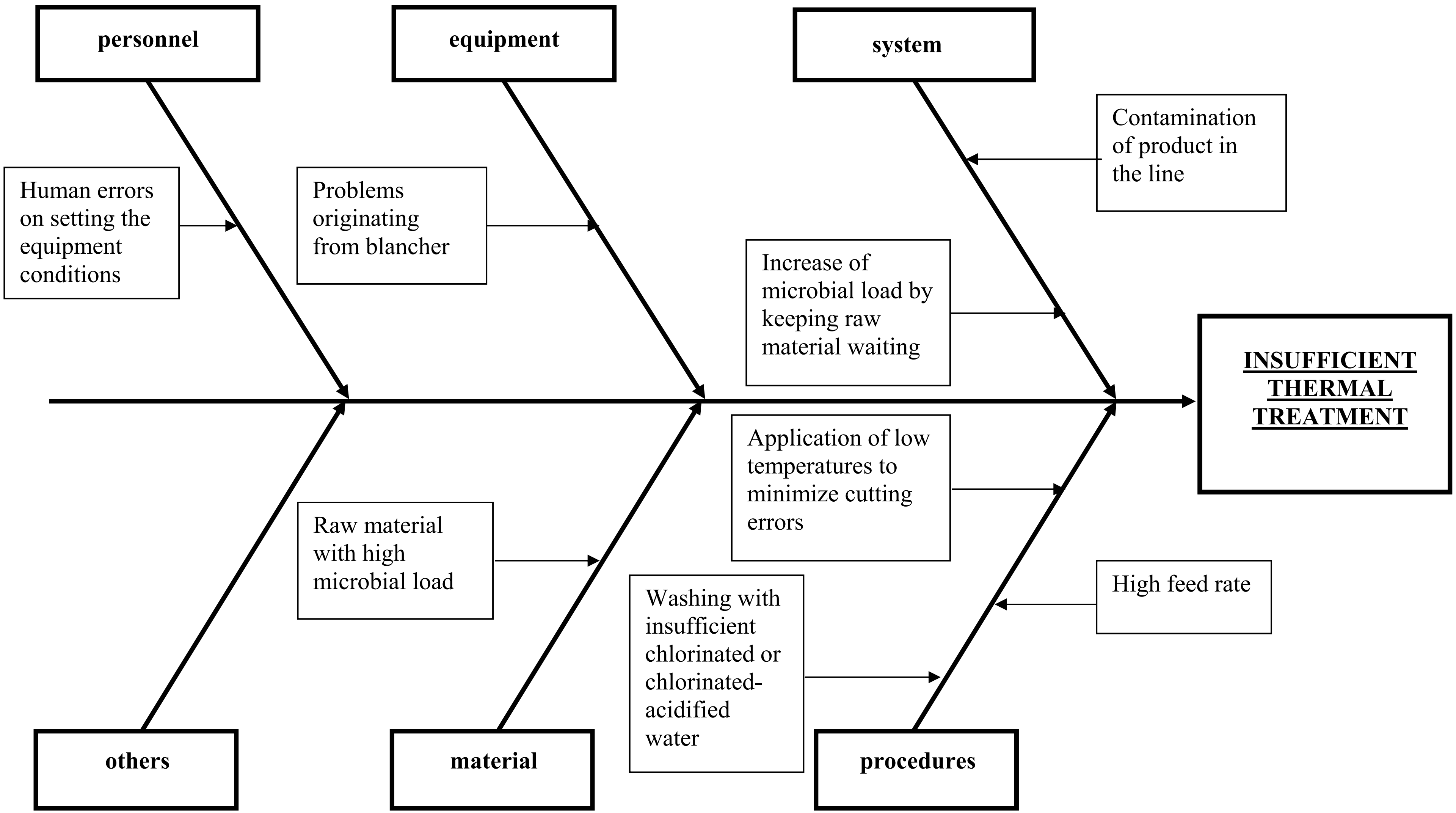

personnel
equipment
system
Human errors on setting the equipment conditions
Problems originating from blancher
Contamination of product in the line
Increase of microbial load by keeping raw material waiting
INSUFFICIENT THERMAL TREATMENT
Application of low temperatures to minimize cutting errors
Raw material with high microbial load
High feed rate
Washing with insufficient chlorinated or chlorinated-acidified water
others
material
procedures

Figure 5.29. Cause and Effect Diagram of CCP-1

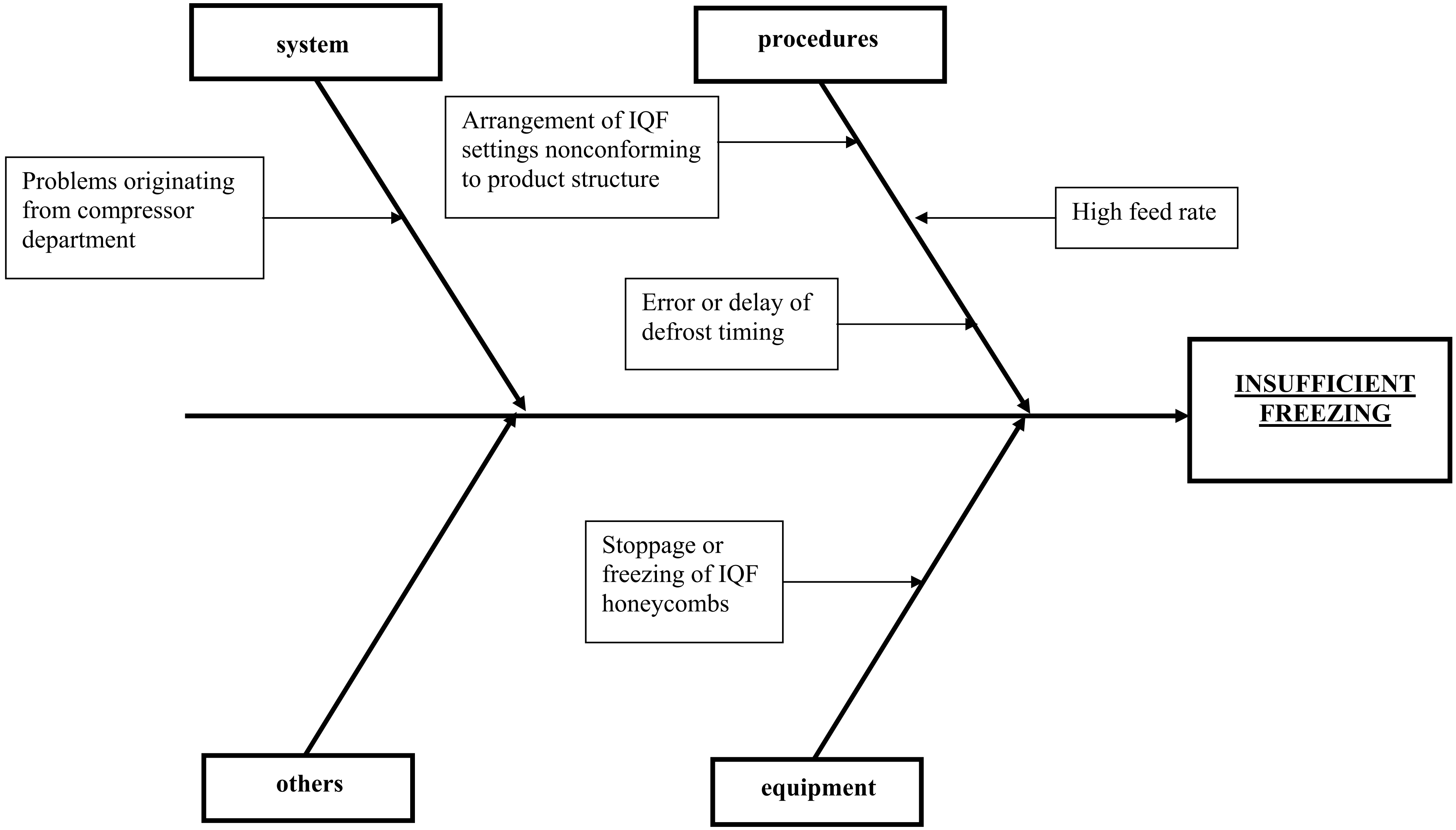

system
procedures
Problems originating from compressor department
Arrangement of IQF settings nonconforming to product structure
High feed rate
Error or delay of defrost timing
INSUFFICIENT FREEZING
Stoppage or freezing of IQF honeycombs
others
equipment

Figure 5.30. Cause and Effect Diagram of CCP-2

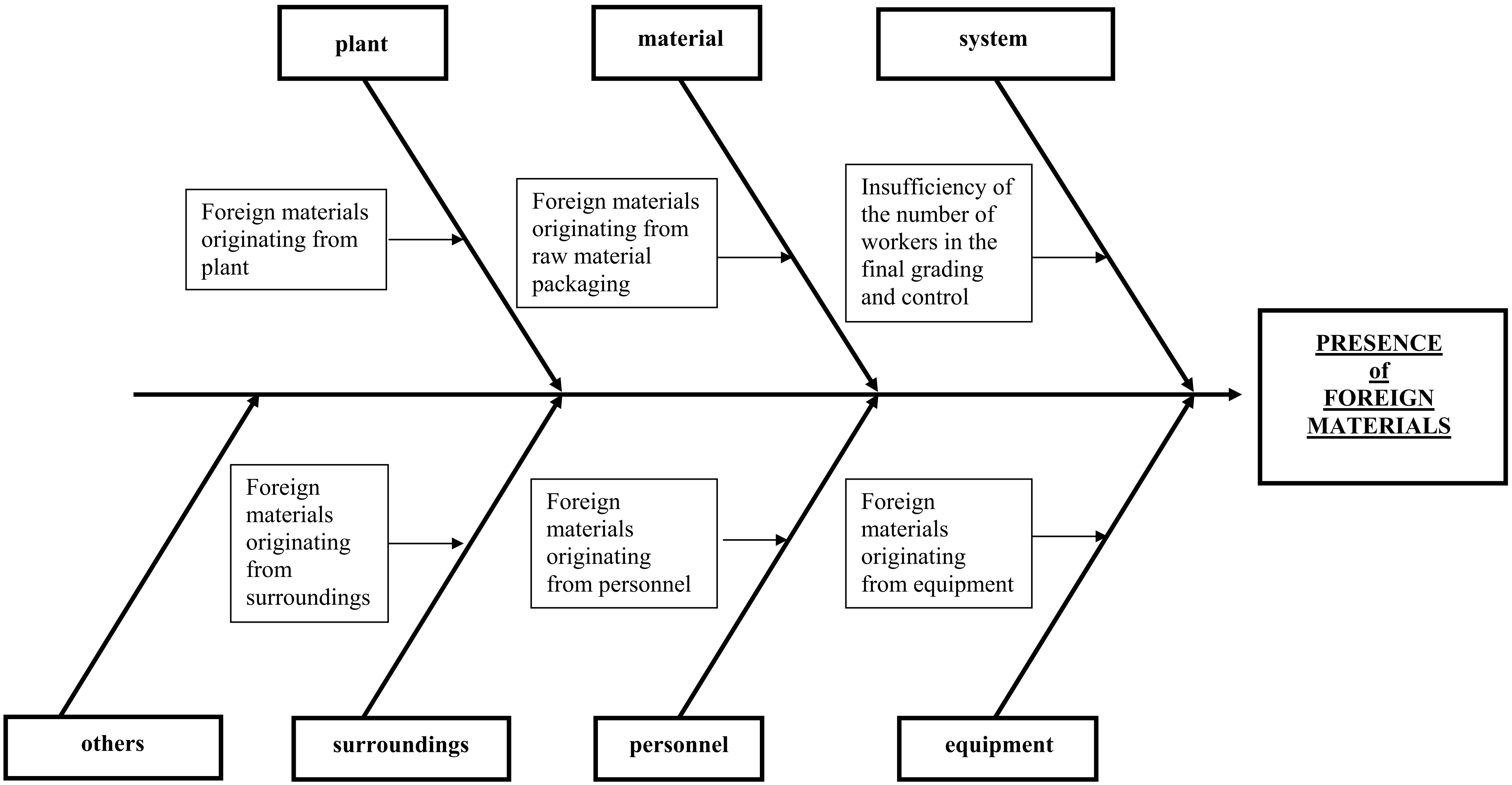

plant
material
system
Foreign materials originating from plant
Foreign materials originating from raw material packaging
Insufficiency of the number of workers in the final grading and control
PRESENCE of FOREIGN MATERIALS
Foreign materials originating from surroundings
Foreign materials originating from personnel
Foreign materials originating from equipment
others
surroundings
personnel
equipment

Figure 5.31. Cause and Effect Diagram of CCP-3

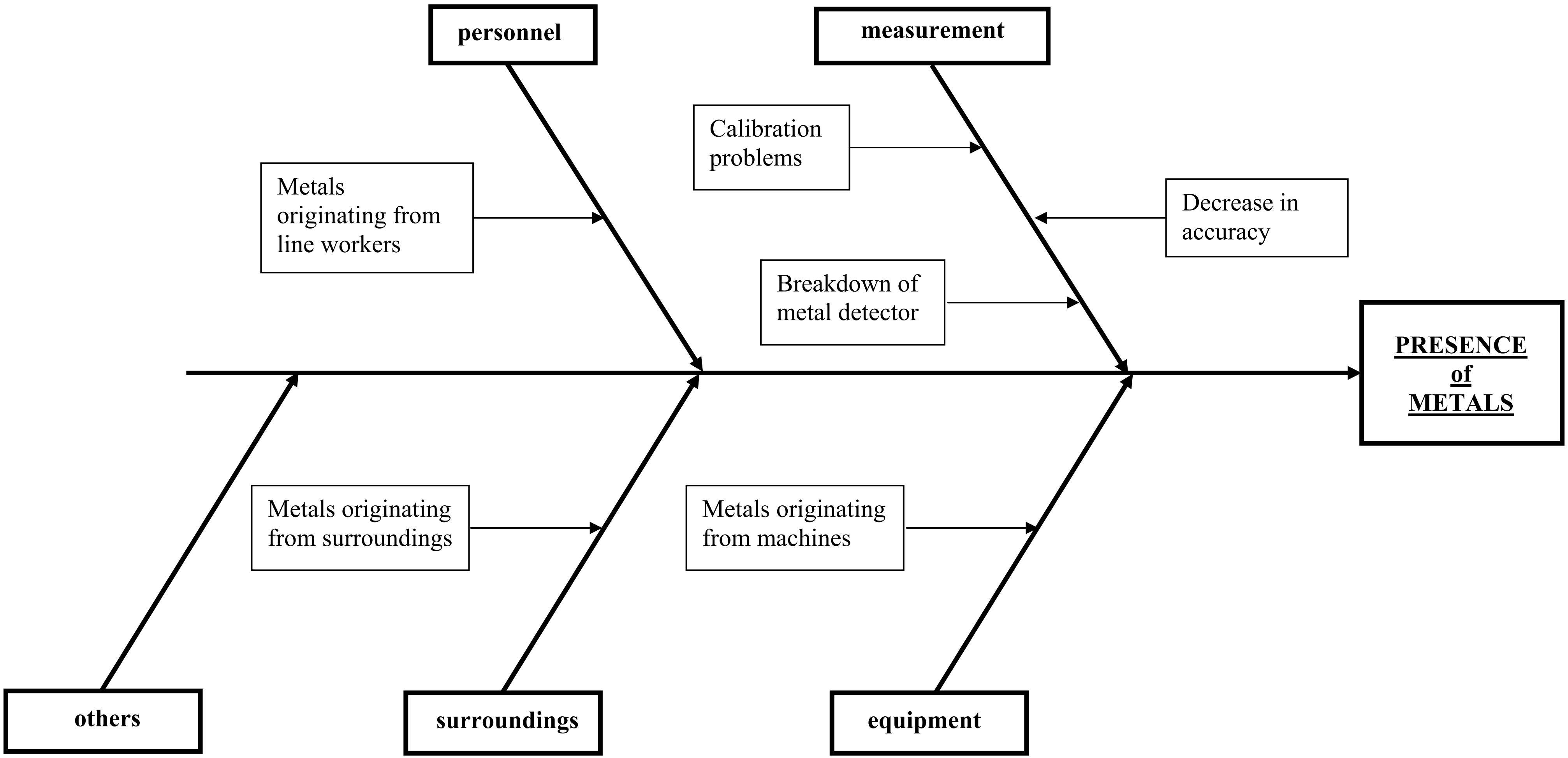

personnel
measurement
Calibration problems
Metals originating from line workers
Decrease in accuracy
Breakdown of metal detector
PRESENCE of METALS
Metals originating from surroundings
Metals originating from machines
others
surroundings
equipment

Figure 5.32. Cause and Effect Diagram of CCP-4

Chapter 6
CONCLUSIONS AND RECOMMENDATIONS

Process and quality data of a frozen food production line were statistically modeled and monitored by using multivariate statistical process monitoring techniques and tools for effective application of HACCP program of the plant.

The process data were analyzed by principal component analysis to overcome the correlation within the variables by projecting high dimensional data space to a lower dimensional model space. The uncorrelated model outputs which are the score variables were used to monitor the process. The out of control observations were successfully observed in Hotelling's T^2 and SPE charts. The product quality data were also analyzed by principal component analysis both to obtain a link between statistical modeling of product data and HACCP program and to check the correlated structure of product data matrix.

Regression methods were applied to process data and product quality data to build a model for the prediction of quality variables with the present information of process and raw material data. The monitoring procedure was performed with multivariate statistical monitoring charts (T^2 and SPE) that determined almost all out of control observations. However, the prediction goal could not be achieved due to the problems encountered in the selection of in-control data.

It was stated that the proper data collection in the production line would cause an enhancement in the application of multivariate statistical techniques, in both monitoring and prediction of critical control point measurements.

HACCP programs require taking measurements at the critical control points and other steps of the production, where data collection is available. Therefore, massive amount of observations belonging to many process variables accumulates. The efficient evaluation of process data is, therefore, necessary to be able to reveal any malfunction in the production. This is the main purpose of applying HACCP plans in food industries. Otherwise, there will be no use of wasting time and money to take measurement. As a result, the statistical charting techniques are strongly recommended in the HACCP programs.

According to the experience and results gained at the end of this study, the recommendations stated below will definitely improve the efficiency of techniques suggested for HACCP.

- to increase the prediction ability of regression techniques with a data set collected under normal operating conditions.
- to develop a procedure to estimate the missing observations in process data.
- to set up a real time process monitoring system.

REFERENCES

Arıkbay C., Gıda Sektöründe Kalite Yönetim Sistemleri ve HACCP. Milli Prodüktivite Merkezi Yayınları, No: 660, Ankara (2002).

Barendsz A. W. (1998). Food Safety and Total Quality Management. *Food Control*, **9** (2-3), pp. 163-170.

Buco S. M. (1990). How Good are Your Results? An Approach to Qualitative and Quantitative Statistical Analysis for Food Monitoring and Process Control Systems. *Food Control*, **1** (1), pp. 40-46.

Conlin A. K., Martin E. B., Morris A. J. (2000). Confidence Limits for Contribution Plots. *Journal of Chemometrics*, **14**, pp. 725-736.

Çınar A., Balasubramaniam V. M., DeCicco J., Martino C., Verdoorn R. (1999). paper [63h]-Multivariate Statistical Monitoring of Cooked Sausage Processes. 1999 Annual Meeting of AIChE, presented at [63]-Sensors and Control in Food Processing.

Does R. J. M. M., Roes K. C. B., Trip A. (1999). Handling Multivariate Problems with Univariate Control Charts. *Journal of Chemometrics*, **13**, pp. 353-369.

Ehiri J. E., Morris G. P., McEwen J., (1995). Implementation of HACCP in Food Businesses: the Way Ahead. *Food Control*, **6** (6), pp. 341-345.

Geladi P. and Kowalski B. R. (1986). Partial Least-Squares Regression: a Tutorial. *Analytica Chimica Acta*, **185**, pp. 1-17.

Gonzales G., Mendez E. M. P., Sanchez M., (2000). Data Evaluation for Soft Drink Quality Control Using PCA and Back-propagation Neural Networks. *Journal of Food Protection*, **63** (12), pp. 1719-1724.

Gonzalez-Miret M. L., Coello M. T., Alonso S., Heredia F. J. (2001). Validation of Parameters in HACCP Verification Using Univariate and Multivariate Statistics. Application to the Final Phases of Poultry Meat Production. *Food Control*, **12**, pp. 261-268.

Hayes G. D., Scallan A. J., Wong J. H. F. (1997). Applying Statistical Process Control to Monitor and Evaluate the HACCP Hygiene Data. *Food Control*, **8** (4), pp. 173-176.

Johnson R. A. and Wichern D. W., Applied Multivariate Statistical Analysis, 4th ed. Prentice-Hall, Inc. Upper Saddle River (1998).

Jouve J. L., Stringer M. F., Baird-Parker A. C. (1999). Food Safety Management Tools. *Food Science and Technology Today*, **13** (2), pp. 82-91.

Kourti T. and MacGregor J. F. (1995). Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods. *Chemometrics and Intelligent Laboratory Systems*, **28**, pp. 3-21.

Kösebalaban F. and Çınar A. (2001). Integration of Multivariate SPM and FDD by Parity Space Technique for a Food Pasteurization Process. *Computers and Chemical Engineering*, **25**, pp. 473-491.

Kresta J. V., MacGregor J. F., Marlin T. E. (1991). Multivariate Statistical Monitoring of Process Operating Performance. *The Canadian Journal of Chemical Engineering*, **69**, pp. 35-47.

MacGregor J. F., Jaeckle C., Kiparissides C., Koutoudi M. (1994). Process Monitoring and Diagnosis by Multiblock PLS Methods. *Process Systems Engineering*, **40** (5), pp. 826-838.

Maesschalck R., Estienne F., Verdú-Andrés J., Candolfi A., Centner V., Despagne F., Jouan-Rimbaud D., Walczak B., Massart D. L., Jong S., Noord O. E., Puel C., Vandeginste B. M. G. (1999). The Development of Calibration Models for Spectroscopic Data Using PCR. *Internet Journal of Chemistry*, **2** (19), [ISSN: 1099-8292].

Martens H. and Russwurm H., Food Research and Data Analysis. Applied Science Publishers, London (1983).

Martin E. B., Morris A. J., Kiparissides C. (1999). Manufacturing Performance Enhancement Through Multivariate Statistical Process Control. *Annual Reviews in Control*, **23**, pp. 35-44.

Mason R. L. and Young J. C. (2000). Interpretive Features of a T^2 Chart In Multivariate SPC. *Quality Progress*, **4**, pp. 84-89.

Mellinger M. (1987). Multivariate Data Analysis: Its Methods. *Chemometrics and Intelligent Laboratory Systems*, **2**, pp. 29-36.

Miller P., Swanson R. E., Heckler J. F. (1998). Contribution Plots: a Missing Link in Multivariate Quality Control. *International Journal of Applied Mathematics and Computer Science*, **8** (4), pp. 775-792.

Miller T. and Balch B. (1991). Statistical Process Control in Food Processing. *ISA Transactions*, **30** (1), pp. 35-37.

Montgomery D. C., Introduction to Statistical Quality Control, 4th ed. John Wiley & Sons, Inc. New York (2000).

Mortimore S. and Wallace C., HACCP a Practical Approach. Aspen Publishers, Inc. Gaithersburg, Maryland (1998).

Negiz A., Ramanauskas P., Çınar A., Schlesser J. E., Armstrong D. J. (1998). Modeling, Monitoring and Control Strategies for High Temperature Short Time Pasteurization Systems. *Food Control*, **9** (1), pp. 1-48.

Nijhuis A., Jong S., Vandeginste B. G. M. (1997). Multivariate Statistical Process Control in Chromatography. *Chemometrics and Intelligent Laboratory Systems*, **38**, pp. 51-62.

Resmi Gazete 24937 (2002). Gıdaların Üretimi, Tüketimi ve Denetlenmesine Dair Yönetmelik, 4. Bölüm Kontrol ve Denetim, Madde 9.

Rodriguez R. N. and Tobias R. D. (1999). Multivariate Methods for Process Knowledge Discovery: The Power to Know Your Process. *Statistical Data Analysis and Data Mining*. Paper 252-26.

Runger G. C. and Montgomery D. C. (1997). Multivariate and Univariate Process Control: Geometry and Shift Directions. *Quality and Reliability Engineering International*, **13**, pp. 153-158.

Sahni N. S., Eide O., Naes T. (1999). An Application of Multivariate Analysis in Product Development in the Food Industry. *Quality Engineering*, **11** (4), pp. 579-586.

Schaffner D. W. (1998). Predictive Food Microbiology Gedanken Experiment: Why Do Microbial Growth Data Require a Transformation? *Food Microbiology*, **15**, pp. 185-189.

Srikaeo K. and Hourigan J. A., (2002). The Use of Statistical Process Control (SPC) to Enhance the Validation of Critical Control Points (CCPs) in Shell Egg Washing. *Food Control*, **13**, pp. 263-273.

Surak J. G., Cawley J. L., Hussain S. A. (1998). Integrating HACCP and SPC. *Food Quality*, **5** (4), p. 46.

Topal R. Ş., Gıda Endüstrisinde Risk Yönetimi Sistemi: HACCP ve Uygulamaları. Yıldız Teknik Üniversitesi, İstanbul (2001).

Westerhuis J. A., Gurden S. P., Smilde A. K. (2000). Generalized Contribution Plots in Multivariate Statistical Process Monitoring. *Chemometrics and Intelligent Laboratory Systems*, **51**, pp. 95-114.

Wikström C., Albano C., Erikkson L., Friden H., Johansson E., Nordahl A., Rannar S., Sandberg M., Kettaneh-Wold N., Wold S. (1998). Multivariate Process and Quality Monitoring Applied to an Electrolysis Process, Part I. Process Supervision with Multivariate Control Charts. *Chemometrics and Intelligent Laboratory Systems*, **42**, pp. 221-231.

Wold S., Sjöström M., Erikkson L. (2001). PLS-Regression: a Basic Tool of Chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58**, pp. 109-130.

APPENDICES

APPENDIX A1
PCA Algorithm

Transform and scale the data and in-control data matrices of **X**. Then, apply the following procedure.

In-control data set modeling ;

- Find eigenvalues (λ) and eigenvectors (**V**) of covariance matrix of in-control data set.
- Set in-control loading matrix (**P**) equal to the eigenvector matrix (**V**).
- Draw the scree plot which is the plot of eigenvalues (λ) from the largest to the smallest and look for an elbow (bend) in the scree plot.
- Calculate percentage of total variance explained by each principal component (columns of **P** matrix) by dividing corresponding eigenvalue to sum of eigenvalues $\left(\frac{\lambda i}{\sum \lambda i}\right)$.
- Find the cumulative sum of percentages of the total variance and plot.
- Decide how many principal components (r) are desired for in-control data set modeling. Use scree plot and cumulative percentage of the total variance for this decision.
- Calculate in-control score vectors (**t**) from 1 to r by using in-control data (**X**) and in-control loading vectors $(\mathbf{p})$: $\mathbf{t} = \mathbf{Xp}$
- Calculate in-control errors (**E**) by using in-control data (**X**), in-control scores (**T**) and transpose of columns of in-control loading matrix from 1 to r (**P'**): $\mathbf{E} = \mathbf{X} - \mathbf{TP'}$

Data set modeling ;

- Calculate score vectors (**t**) from 1 to r by using data (**X**) and in-control loading vectors $(\mathbf{p})$: $\mathbf{t} = \mathbf{Xp}$
- Calculate errors (**E**) by using data (**X**), scores (**T**) and columns of transpose of loadings (**P'**) from 1 to r : $\mathbf{E} = \mathbf{X} - \mathbf{TP'}$

APPENDIX A2
PCR Algorithm

Transform and scale the data and in-control data matrices of **X** and **Y**. Then, apply the following procedure.

In-control data set modeling ;

- Find eigenvalues and eigenvectors of covariance matrix of in-control data set of **X** by standard value decomposition.
- Equalize the eigenvector matrix to in-control loading matrix of **X** (**P**).
- Draw the scree plot which is the plot of eigenvalues (λ) from largest to smallest and look for an elbow (bend) in the scree plot.
- Calculate percentage of the total variance by dividing each eigenvalue to sum of eigenvalues and cumulative sum of the percentages.
- Plot cumulative percentage of the total variance.
- Decide how many principal components (r) are desired for in-control data set modeling. Use scree plot and cumulative percentage of the total variance for this decision.
- Calculate in-control score vectors of **X** (**t**) from 1 to r by using in-control data (**X**) and in-control loading vectors $(\mathbf{p})$: $\mathbf{t} = \mathbf{X}\mathbf{p}$
- Calculate in-control errors of **X** (**E**) by using in-control data (**X**), in-control scores (**T**) and transpose of columns of in-control loading matrix from 1 to r (**P'**) : $\mathbf{E} = \mathbf{X} - \mathbf{T}\mathbf{P}'$
- Calculate least square estimate of regression coefficients matrix (β) by in-control scores of $\mathbf{X}\,(\mathbf{T})$ and in-control $\mathbf{Y}$: $\boldsymbol{\beta} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y}$
- Calculate in-control errors of **Y** (**F**) by in-control **Y**, in-control scores of **X** (**T**) and β : $\mathbf{Y} = \mathbf{T}\boldsymbol{\beta} + \mathbf{F}$

Data set modeling ;

- Calculate score vectors of **X** (**t**) from 1 to r by using data (**X**) and in-control loading vectors $(\mathbf{p})$: $\mathbf{t} = \mathbf{X}\mathbf{p}$
- Calculate errors (**F**) by using data (**Y**), scores (**T**) and β : $\mathbf{F} = \mathbf{Y} - \mathbf{T}\boldsymbol{\beta}$

APPENDIX A3

PLSR Algorithm (NIPALS Algorithm)

Transform and scale the data and in-control data matrices of **X** and **Y**. Then, apply the following procedure.

- Start : Set **u** equal to any column of **Y**.
- Regress columns of **X** on **u** to get **X** weights : $\mathbf{w'} = \mathbf{u'X/u'u}$
- Normalize **w** to unit length.
- Calculate **X** scores : $\mathbf{t} = \mathbf{Xw/w'w}$
- Regress to columns of **Y** on **t** to get **Y** loadings : $\mathbf{q'} = \mathbf{t'Y/t't}$
- Normalize **q** to unit length.
- Calculate the new score vector for **Y** : $\mathbf{u} = \mathbf{Yq/q'q}$
- Check convergence of **u** : compare **t** with the one from the preceding iteration. If they are equal (within a certain rounding error) go to the next step, else go to the step which calculates **X** weights.
- Calculate **X** loadings by regressing columns of **X** on **t** : $\mathbf{p'} = \mathbf{t'X/t't}$
- Find the regression coefficient **b** for the inner relation : $\mathbf{b} = \mathbf{u't/t't}$
- Calculate residual matrices for the outer relation : $\mathbf{E} = \mathbf{X} - \mathbf{tp'}$; $\mathbf{F} = \mathbf{Y} - \mathbf{btq'}$
- To calculate the next set of latent vectors replace **X** and **Y** by **E** and **F** and repeat.

(Geladi and Kowalski, 1986; MacGregor *et al*., 1994)

APPENDIX B1

In-control Data Set

Table B1.1. In-control Data Set

7.1	4.6	6.0	4.4	0.0	7.4	70	92	2	22000	0	240	100
9.3	5.7	5.8	4.6	0.8	7.8	70	92	5	34000	240	1200	200
12.0	5.3	4.2	3.8	0.0	7.9	70	92	4	34000	240	1200	200
1.2	6.6	6.0	4.5	1.4	7.2	70	91	4	25500	10	100	100
0.0	4.0	5.4	4.6	0.6	7.0	70	92	4	19500	0	100	100
7.8	9.3	4.6	3.6	0.3	6.8	70	92	0	292500	0	3000	100
10.5	12.7	8.9	5.4	0.5	6.6	65	93	4	19500	60	100	100
12.9	10.6	5.3	11.9	4.8	6.5	65	93	1	19500	60	100	100
11.4	12.6	10.5	11.1	3.2	7.6	65	92	1	19500	60	100	100
9.5	4.4	7.7	7.1	0.0	7.7	65	91	0	6000	40	100	100
10.1	7.1	8.9	7.5	2.3	7.6	65	92	4	38500	0	550	100
4.5	7.0	7.5	6.5	3.2	7.6	65	92	2	38500	0	550	100
8.6	5.5	6.3	7.2	4.7	8.1	65	92	2	18000	160	100	100
7.3	6.4	5.4	6.1	1.3	7.5	65	92	2	15500	70	100	100
3.6	5.2	7.2	6.2	0.0	7.1	65	92	0	23000	50	100	100
9.9	12.0	6.1	6.6	0.0	6.9	70	93	2	37500	40	350	100
7.8	9.7	8.7	7.3	2.7	7.1	70	93	8	37500	40	350	100
7.0	11.3	10.1	8.5	3.9	6.9	65	92	2	20000	50	100	100
4.2	9.8	8.7	7.8	3.9	6.8	65	92	3	20000	50	100	100
10.2	8.2	7.0	7.5	2.0	7.5	70	93	3	35500	360	120	100
10.2	8.9	6.8	8.1	1.4	7.2	65	92	0	78500	10	600	100
14.0	4.9	7.2	8.0	0.0	7.2	65	92	2	78500	10	600	100
13.9	9.3	8.1	4.7	6.8	7.5	60	94	2	11500	0	2700	200
9.8	7.9	8.3	7.2	2.5	5.7	65	92	1	2000	70	100	100

APPENDIX B2

Descriptive Statistics of In-control Data Set

Table B2.1. Descriptive Statistics of In-control Data Set

		variables	mean	min.	max.	standard deviation	units
X process variables	x1	Plant origin foreign material	8,45	0,00	14,00	3,6575	%w/w
	x2	Rotten	7,88	4,00	12,70	2,7189	%w/w
	x3	Burst	7,11	4,20	10,50	1,6768	%w/w
	x4	Diseased	6,68	3,60	11,90	2,0862	%w/w
	x5	Crumpled	1,93	0,00	6,80	1,8966	%w/w
	x6	Brix	7,22	5,70	8,10	0,5256	°Brix
	x7	Blanching time	66,67	60	70	2,8233	s
	x8	Blanching temperature	92,21	91	94	0,6580	°C
Y product quality variables	y1	Foreign material	2	0	8	2	count/w
	y2	Total viable count	39438	2000	292500	56960,0000	cfu/g
	y3	*E.coli*	68	0	360	91,0992	cfu/g
	y4	Yeast	528	100	3000	787,2310	cfu/g
	y5	Mold	113	100	200	33,7832	cfu/g

APPENDIX C1

Model Parameters of PCA Analysis

PCA Analysis of Process Data (X) :

Number of principal components (PC's) : 4

Total variability explained by principal components : 80.99%

Table C1.1. Variability Explained by Principle Components in PCA Model of **X** Data

Principal components	Variability explained by principal components
PC 1	40.21%
PC 2	15.84%
PC 3	15.19%
PC 4	9.75%

Loading matrix of the model (**P**) :

$$\begin{array}{cc} & \begin{array}{rrrr} \text{PC 1} & \text{PC 2} & \text{PC 3} & \text{PC 4} \end{array} \\ \mathbf{P} = & \left[\begin{array}{rrrr} 0.2921 & -0.0711 & -0.5720 & -0.5666 \\ 0.4137 & 0.4341 & 0.0391 & -0.0485 \\ 0.3833 & -0.1429 & 0.4166 & -0.0348 \\ 0.3905 & -0.1309 & 0.2479 & -0.4879 \\ 0.4069 & -0.2460 & -0.0421 & 0.5568 \\ -0.1640 & -0.6514 & -0.3578 & -0.0102 \\ -0.3832 & 0.4262 & -0.0789 & -0.1456 \\ 0.3238 & 0.3202 & -0.5480 & 0.3256 \end{array}\right] \end{array}$$

PCA Analysis of Product Data (Y) :

Number of principal components : 3

Total variability explained by principal components : 86.44%

Table C1.2. Variability Explained by Principle Components in PCA Model of **Y** Data

Principal components	Variability explained by principal components
PC 1	40.10%
PC 2	29.35%
PC 3	16.99%

Loading matrix of the model (**P**) :

$$\begin{array}{cc} & \begin{array}{ccc} \text{PC 1} & \text{PC 2} & \text{PC 3} \end{array} \\ \mathbf{P} = & \left[\begin{array}{ccc} 0.1833 & -0.5273 & 0.6588 \\ -0.5356 & 0.2627 & -0.1223 \\ 0.3867 & -0.3212 & -0.7140 \\ -0.6667 & -0.2334 & -0.0481 \\ -0.2925 & -0.7038 & -0.1975 \end{array}\right] \end{array}$$

APPENDIX C2

Model Parameters of MLR Analysis

Regression coefficients matrix of the model (β_{MLR}) :

$$\boldsymbol{\beta} = \begin{bmatrix} -0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0713 & 0.0471 & 0.2657 & 0.5650 & 0.5515 \\ -0.2965 & 0.5409 & 0.0848 & 0.1985 & -0.0588 \\ 0.4422 & -0.3846 & -0.0284 & -0.1621 & -0.0349 \\ -0.1290 & -0.0748 & 0.3394 & -0.7351 & -0.6717 \\ 0.5195 & -0.2383 & -0.0254 & 0.2583 & 0.4287 \\ 0.0263 & 0.2387 & 0.0920 & 0.1019 & 0.1806 \\ 0.7217 & 0.0946 & 0.3082 & 0.0252 & 0.1179 \\ 0.2954 & -0.0914 & -0.1089 & -0.0917 & -0.0235 \end{bmatrix}$$

APPENDIX C3

Model Parameters of PCR Analysis

Number of principal components : 4

Total variability explained by principal components : 80.99%

Table C3.1. Variability Explained by Principle Components in PCR Model

Principal components	Variability explained by principal components
PC 1	40.21%
PC 2	15.84%
PC 3	15.19%
PC 4	9.75%

Loading matrix of the model (**P**) :

$$\begin{array}{c} \\ \mathbf{P} = \end{array}\begin{array}{cccc} \text{PC 1} & \text{PC 2} & \text{PC 3} & \text{PC 4} \\ \end{array}$$

$$\mathbf{P} = \begin{bmatrix} 0.2921 & -0.0711 & -0.5720 & -0.5666 \\ 0.4137 & 0.4341 & 0.0391 & -0.0485 \\ 0.3833 & -0.1429 & 0.4166 & -0.0348 \\ 0.3905 & -0.1309 & 0.2479 & -0.4879 \\ 0.4069 & -0.2460 & -0.0421 & 0.5568 \\ -0.1640 & -0.6514 & -0.3578 & -0.0102 \\ -0.3832 & 0.4262 & -0.0789 & -0.1456 \\ 0.3238 & 0.3202 & -0.5480 & 0.3256 \end{bmatrix}$$

Regression coefficients matrix of the model (β_{PCR}) :

$$\boldsymbol{\beta} = \begin{bmatrix} 0.0434 & -0.1411 & 0.0555 & -0.0531 & -0.0470 \\ 0.0772 & 0.2104 & 0.0203 & 0.0168 & -0.1523 \\ -0.1503 & -0.2173 & -0.0728 & -0.5642 & -0.5779 \\ 0.3015 & -0.1817 & -0.4147 & 0.1438 & 0.2313 \end{bmatrix}$$

APPENDIX C4

Model Parameters of PLSR Analysis

Number of latent variables (LV's) of **X** : 4

Number of latent variables of **Y** : 4

Total variability of **X** explained by latent variables: %76.12

Total variability of **Y** explained by latent variables: %34.82

Table C4.1. Variability Explained by Latent Variables in PLSR Model

X block		**Y** block	
Latent variables of **X**	Variability of **X** explained by latent variables	Latent variables of **Y**	Variability of **Y** explained by latent variables
LV 1	20.45%	LV 1	19.34%
LV 2	33.01%	LV 2	4.74%
LV 3	12.61%	LV 3	5.73%
LV 4	10.04%	LV 4	5.02%

Loading matrix of **X** matrix (**P**) :

$$\begin{array}{cc} & \begin{array}{cccc} \text{LV 1} & \text{LV 2} & \text{LV 3} & \text{LV 4} \end{array} \\ \mathbf{P} = & \left[\begin{array}{cccc} -0.1959 & 0.3855 & 0.4626 & -0.5749 \\ 0.2899 & 0.3259 & 0.5399 & 0.3305 \\ 0.5261 & 0.2949 & -0.0701 & 0.1032 \\ 0.5289 & 0.2199 & 0.4088 & -0.2659 \\ 0.1982 & 0.4856 & -0.2402 & 0.2292 \\ -0.3935 & 0.0253 & -0.2044 & -0.1603 \\ -0.2880 & -0.4081 & 0.3917 & 0.5261 \\ -0.2097 & 0.4548 & 0.2639 & 0.3520 \end{array}\right] \end{array}$$

Loading matrix of **Y** matrix (**Q**) :

$$\mathbf{Q} = \begin{array}{c} \begin{array}{cccc} \text{LV 1} & \text{LV 2} & \text{LV 3} & \text{LV 4} \end{array} \\ \begin{bmatrix} -0.1466 & 0.4196 & 0.0371 & 0.9657 \\ -0.3168 & -0.4183 & 0.6224 & 0.1012 \\ 0.0389 & 0.0367 & 0.7295 & -0.2052 \\ -0.6557 & 0.4977 & -0.0333 & -0.1014 \\ -0.6683 & 0.6324 & -0.2790 & -0.0689 \end{bmatrix} \end{array}$$

Weight matrix of the model (**W**) :

$$\mathbf{W} = \begin{bmatrix} -0.4580 & 0.3172 & 0.4639 & -0.5522 \\ 0.0952 & 0.1997 & 0.5854 & 0.1071 \\ 0.4414 & 0.4801 & -0.0927 & 0.2511 \\ 0.6574 & 0.0226 & 0.5389 & -0.1623 \\ -0.0221 & 0.6885 & -0.1898 & 0.3068 \\ -0.4417 & 0.0242 & 0.0472 & -0.0030 \\ -0.1761 & -0.4008 & 0.4122 & 0.7412 \\ -0.3939 & 0.3685 & 0.1711 & 0.3332 \end{bmatrix}$$

Regression coefficients matrix of the model (β_{PLSR}) :

$$\boldsymbol{\beta} = \begin{bmatrix} 0.7689 & 0.2994 & 0.5328 & 0.5587 \end{bmatrix}$$