MULTIVARIATE STATISTICAL ANALYSIS OF DATA FROM AN INDUSTRIAL FLUIDIZED CATALYTIC CRACKING PROCESS USING PCA AND PLS

BY

CAROL FRANCES SLAMA,
B.Sc.

McMASTER UNIVERSITY LIBRARY
TP 690.4 .S53 1991
Multivariate statistical analy C.2
3 9005 0475 6428 5

Thode
TP
690.4
.S53
1991
c.2

ANALYSIS OF INDUSTRIAL FCCU DATA USING PCA AND PLS

MULTIVARIATE STATISTICAL ANALYSIS OF DATA FROM AN INDUSTRIAL FLUIDIZED CATALYTIC CRACKING PROCESS USING PCA AND PLS

By

CAROL FRANCES SLAMA, B. Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Engineering

McMaster University

(c) Copyright by Carol Frances Slama, November 1991

MASTER OF ENGINEERING (1988) McMASTER UNIVERSITY
(Chemical) Hamilton, Ontario

TITLE: Multivariate Statistical Analysis of Data from an Industrial Fluidized Catalytic Cracking Process Using PCA and PLS

AUTHOR: Carol Frances Slama, B.Sc. (University of Ottawa)

SUPERVISOR: Professor J.F. MacGregor

NUMBER OF PAGES: xi, 180

ABSTRACT

Principal Components Analysis (PCA) and Partial Least Squares (PLS, or Projection to Latent Structures) were used to evaluate the process history of a fluidized catalytic cracking unit (FCCU). Specifically, the goals of the work were to identify interesting periods in the process history, identify relationships amongst process variables, develop a predictive model of the product yields and selectivities, and to create a monitoring space to detect process changes and disturbances.

Major process changes of feed rate, feed quality and production modes were easily modelled by the first few latent variables (LVs) in both the PCA and PLS analyses. Later LVs highlighted transients obvious to operations. Plots of the process behaviour in the space of these latent variables were able to clearly reveal where major changes occurred in the process, implying that this approach is useful for the post analysis of historical data bases. Diagnosing the reasons for changes, however, was much more difficult.

PLS was quite successful in obtaining predictive models for the product yields and selectivities. A linear model of eleven dimensions was able to predict 81.3% of the cross-validated sum of squares in the Y space and 78.3% of the sum of squares in the X space. The hierarchical PLS approach of Wold et al. (1987) was also applied to the data set and generated results of similar predictive ability and interpretation.

The development of multivariate SPC monitoring procedures was less successful, due to the FCCU's continually shifting process operations. This latter use of PCA and PLS would be much more amenable to processes with a stable operating point (such as would be found in a quality control situation).

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge the support and vision of my supervisor, Dr. J.F. MacGregor, and the financial support of Shell Canada Limited without whom the successful completion of this internship thesis would not have been possible.

There are many people at Shell whom I would like to thank for their assistance with the collection of the industrial data: Pierre Vadnais, Dave Onderwater, Dominique Jutras, Tom Chupick, Gilles Berube, Pierre Brisson, ICPO, and especially Christian Houle for his help in the analysis phase of this work.

Special thanks must also go to Bert Skagerberg, Jim Kresta and Paul Gossen for their help in unravelling the mysteries of PCA and PLS. Credit is also due to Dr. T.E. Marlin for his assistance with an FCCU simulation (which did not get incorporated into this work).

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure **Page**

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Overview

If one has ever had the opportunity to walk through a modern control room of a typical petrochemical process, one cannot help but be awestruck by the vast amount of data being collected, stored and displayed from process sensors on a minute, hourly and daily basis. The sheer volume of these observations means that only a small percentage of them is ever actually examined by an operator or engineer, with little or no analysis being performed. Intuitively, it is felt that this data should provide some insight about the process since it is the closest "picture" one can get, but little, as yet, has been done with such data and this situation is typical of many process industries.

The work of this thesis is a first attempt at testing the applicability of Principal Components Analysis (PCA) and Partial Least Squares or Projections to Latent Structures (PLS) for analysis and modelling of a large steady-state industrial data set. PCA and PLS appear to have many advantages over traditional multivariate techniques for analyzing process data in that they are capable of dealing with noisy, highly correlated data sets where the number of variables may greatly outweigh the number of samples available. They are also straightforward to use and build distribution-free empirical models where causality is based on correlation (Geladi 1988). Dimensionality reduction is a key aspect of both PCA and PLS; "patterns" or models of the data can be viewed in easy-to-comprehend two dimensional plots or "windows".

The specific goals of the thesis are: i) identification of interesting time periods in the process history or interesting relationships amongst the process variables being collected, ii) development of predictive models of some desirable output or phenomena (i.e., product yields, quality variables) from the operating conditions, and iii) development of a fault detection or monitoring system which provides an indication of process performance (e.g., normal, abnormal), by signalling undesirable changes (faults) in the process and which might aid in assigning a cause or causes to such changes.

1.2 PCA and PLS

PCA is a pattern recognition technique that works with a single data matrix and searches for a small number of latent variables (LVs), linear combinations of the measured variables, which correspond to the directions of greatest variability in the data set.

PLS is a calibration technique used to build models between two blocks of data. Its latent variables are partly summarizing and partly correlation maximizing as they attempt to not only describe variations within the individual blocks but also the correlation structure between the two blocks.

Hierarchical PLS is an extension of the PLS method which allows for more than two blocks of data to be used in the analysis. All three techniques are discussed in detail in Chapter 4.

The latent variables have attractive orthogonality and eigenvector properties which ensure convergence and allow one to summarize the LVs into biased regression models (Wold et al., 1987b). They are used to form low (two and three) dimensional spaces for viewing the relationships amongst samples and variables. They summarize the dominating trends in the data and act as the model for predicting process states.

1.3 Research Approach

The industrial process selected for study was a fluidized catalytic cracking unit (FCCU) located at a Shell Canada Limited petroleum refinery. Originally, it was felt that due to the complexity and high product value of this unit, analyses which might even slightly increase the knowledge-base of the FCCU would be of benefit.

After studying the results, it was found that the operational policies were such that the FCC process never operated long at any one condition, and hence it was difficult to isolate special events or detect faults in such a changing environment. However, some benefits did arise from the PCA and PLS analyses.

An initial data set of approximately three and one half months' worth of hourly averages from over 300 variables was collected. Due to computational restrictions, however, this was pared down to 142 variables and approximately 1400 samples from which specific subsets were used, based on the purpose of the analyses. Table 1.1 outlines the objectives of this thesis.

Data covering the range of operating conditions and time history was used for the first PC analysis, followed by analysis of a subset of data covering approximately eleven days, to illustrate how the samples and variables group in the model space. Predictive models of percentage volume yields and selectivities for the products were built using PLS on the full range of operating data. The statistical process control (SPC) monitoring space was also built using PLS but required a subset of reference data representing "normal" operating data from which to develop the detection "rules" and an additional set of data containing abnormalities with which to test the model. The hierarchical PLS algorithm was tested on the PLS prediction data to examine how the model resolution and results differed when the X space was split into a number of blocks.

Table 1.1--Analyses Performed on Industrial Data

Purpose	Goal	Data Set	Technique
Preliminary Analysis	Find interesting periods in the samples and relationships amongst variables in the data.	Use the whole data to illustrate this concept.	PCA separately on X and Y
	Examine a smaller segment of data for further details.	Use a subset of the above data.	Same as above
Predictive Model	Predict percentage volume yields and selectivities from operating conditions.	Cover as wide a range of operating conditions as possible (same data set as used for first PCA analysis).	PLS
Multivariate SPC Monitoring	Build a monitoring space able to differentiate between "normal" and "abnormal" operations and if possible, identify the cause(s) of abnormalities.	Reference set represents "normal" operations. Test data contains known faults or process changes to be detected.	PLS
Effect of sub-dividing data	Answer the question: how do the model resolution and results differ when data is analyzed using the hierarchical approach?	Use data from PLS (prediction) case; break X data up into six blocks treated as independent of each other.	Hierarchical PLS

1.4 Summary of Chapters to Follow

The remaining body of the thesis is broken down as follows.

Chapter 2 provides a brief literature review of other methods used to handle analysis, modelling and monitoring of industrial processes (with particular attention to FCC units) in order to illustrate how PCA and PLS augment these fields of study. Chapter 3 describes the fluidized catalytic cracking (FCC) process and the typical constraints, disturbances and operational strategies to which it is subjected. The chapter also provides a brief review of FCCU models available and their limitations. Chapter 4 outlines the algorithms for PCA, PLS and hierarchical PLS followed by sections on data pretreatment issues and interpretation tools. Chapters 5, 6, 7 and 8 contain the results

from industrial cases involving preliminary analysis, prediction model building, statistical process monitoring, and the hierarchical case, respectively. Chapter 9 summarizes the conclusions and areas for further work.

CHAPTER 2: LITERATURE REVIEW

This chapter provides a literature review of work done in the areas of analysis, modelling and monitoring of industrial processes (with particular attention to FCC units, where warranted). This should help to illustrate how PCA and PLS augment these fields of study.

2.1 Preliminary Analysis of Process Data

Current on-line software used in industrial data acquisition computers is typically limited in the types of analysis options they make available to the operator or engineer. Most provide trend (or time series) plots indicating the desired set point, mean, minimum or maximum value or perhaps confidence intervals of a variable. Due to the large number of sensors available, these plots are usually reserved for those variables known to have a significant effect on the process and which are probably part of a control strategy or have specifications which warrant monitoring. This part of a data acquisition system is limited not only by the amount of screen display space available but also by the CPU time available for keeping it updated.

More sophisticated approaches to identifying interesting periods in the data have arisen from the discrete parts manufacturing industries. These include the Shewhart, cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) charts.

Although these were designed as part of control monitoring strategies (which involve looking for assignable causes to deviations or changes and correction of the process to minimize reoccurrence) they are used, as a first step, to detect process changes.

All three involve collection of a sample (or group of samples), calculation of an appropriate statistic and then plotting of that statistic on a chart as a function of the sampling sequence or time. The Shewhart chart consists of plotting a mean value and an estimated standard deviation (called a range value), while the CUSUM chart plots information about the distribution of the variable being monitored (usually the sum of its deviations from a target value). The EWMA chart uses information from past samples together with the most recent measurement to determine a smoothed sample mean.

Hypothesis tests on the samples are performed, by means of confidence limits or "masks", to determine whether or not a significant process change has occurred. They also provide an estimate of times in the past when changes occurred, as well as an estimate of the size of the change (and thus the size of corrective action required), and a measure of the quality of an output for classification (Himmelblau 1978).

These graphical means of analysis are easy to implement (no deterministic process model is required) and are simple to maintain and use under plant operating conditions. However, each involves making some assumptions about the statistics of the variable(s) being monitored, namely, that they can be represented as having some fixed target value with constant variance and are independent and identically distributed unless a non-random change has occurred.

Most variables in process industries are being monitored at a much higher frequency than these charts were initially intended. Samples or observations of the process variables are highly correlated in time and this is a direct violation of the independence and identically distributed assumptions. It means that the hypothesis boundaries (in the form of confidence limits or masks) do not have a statistically sound

foundation and can lead to erroneous and/or long delays in detection of process changes. This issue is discussed in papers by MacGregor (1988), Harris and Ross (1991), Johnson and Bagshaw (1974), Kemp (1967) and the text by Himmelblau (1978).

A second issue is that these univariate methods do not use any information about the relationships amongst the variables being charted. Process data can be expected to be highly correlated (due to the large number of redundant variables available for monitoring and the relatively few underlying phenomena taking place in a process at any given time). Pooling and drawing conclusions from the results of individually monitored variables can be dubious.

Some adaptations to statistical tests have been made to accommodate the multivariate case and should be noted.

The Hotelling T^2 test is an extension of the Student t test which takes into account the joint normal distributions of the variables being monitored. Its use is discussed in detail in Hotelling (1947) using the example of quality testing of airplane bombsights, and more recently in Anderson (1984). A multivariate CUSUM chart is also described by Healy (1987).

However, even the use of these methods to data sets containing tens or hundreds of variables can be quite cumbersome and impractical. This spurred the development of multivariate pattern recognition techniques.

2.2 Multivariate Techniques

Pattern recognition techniques are applied to multivariate data sets containing specific classes of samples to find the "typical data pattern" of each class. These "patterns" become classification rules which are then used to assign new samples (subject to the same series of measurements) to one of the known classes (Wold et al.

1983). Many different types of pattern recognition techniques exist and selection of the method to use depends heavily upon the type of data being analyzed, the problem being assessed and the objectives of the analysis.

Cluster analysis is used to check for the existence of groupings in the samples or groupings in the variables. Factor analysis finds the linear combinations of variables that describe the directions of greatest variability in the data. If one wishes to search for the systematic differences between known classes of samples, there are the classification techniques of linear discriminant analysis, K nearest neighbour, and the probability density function (Bayesian) method. Although these techniques appear to offer a wide range of analysis options, several are hampered by restrictions on the number of variables allowed (i.e., must be much less than the number of samples) and the inability to cope with outliers. Details on these techniques are found in Chatfield and Collins (1980), Mardia, Kent and Bibby (1979), Martens, Wold and Martens (1983) and, Wold et al. (1984).

2.3 Predictive Models

The approaches available for predictive model development span from those built on the theoretical principles of conservation and continuity (e.g., mass, energy and momentum equations) to development of regression models based simply on data collected from the process. Each comes with its own set of advantages and limitations.

Several theoretical models have been developed for the catalytic cracking process and these are discussed further in Chapter 3. Their basis in fundamental engineering concepts makes them easy to understand and interpret, allows for a relatively extensive prediction range and they generally yield robust parameters. However, they require long development times and assumptions must be made to simplify the model and cut down on the solution time requirements. If the model is to

represent a real-life process, its parameters will have to be fitted to match plant characteristics (which may require plant experimentation and further assumptions). Non-linear models pose a greater solution challenge than linear ones and as the desired operating range to be modelled increases, so does the complexity of the model (and hence the above problems).

Regression analysis is a quick way to develop a process model from a handful of knowingly important variables. The simplest model structure used is the linear equation. Available data is used to estimate the parameters of the fitted equation. The predictions are then gauged against some criterion and a check of the underlying assumptions is made.

Many methods exist for solving the parameters of such equations (such as the methods of least squares, weighted least squares, maximum likelihood and Bayes), each with its own criterion and requirements concerning a priori information. For model equations linear in the coefficients, the least squares method will yield coefficient estimates which are unbiased if the predicted errors are uncorrelated and have the same probability distribution (Himmelblau 1978). However, if the process variables are highly correlated (as is always the case with undesigned process data) the least squares estimates will have very large variances.

Violations of the assumptions underlying the solution method can generate misleading results. The least squares method assumes the model form is appropriate and that the errors are independent, normally distributed random variables with zero mean and constant variance. Models using variables straight from an industrial process rarely pass these requirements unless the data has been specifically designed to overcome this. Box, Hunter and Hunter (1978) discuss this issue in depth.

Other problems in fitting equations to such data sets include inconsistencies amongst data, presence of controlled variables, serially correlated errors and dynamic relationships, not to mention the danger of drawing causal conclusions from correlational relationships (Box, Hunter and Hunter 1978).

An important class of stochastic models for describing dynamic systems are the autoregressive-moving average processes (ARIMA processes) discussed by Box and Jenkins (1976). No prior information about the model structure is needed; this is checked as the model is built. This method can provide a good representation of a process although the form of the model may be hard to interpret in terms of fundamental process principles.

Neural networks are another class of empirical models. An initial structure (consisting of layers of nodes) is selected and data is used to "train" or build the input-output relationships. Neural networks can accommodate non-linear relationships, but they are also sensitive to data pretreatment methods, selection of appropriate exemplars and inner nodes, long training times and defining adequate training sets.

Clearly, the most important consideration for selection of a predictive model type is the purpose for which the model will be used. Simple empirical models are useful when passive prediction of process outputs is all that is required, while sophisticated mechanistic models are used in many optimization projects (Ramesh and Davis 1989; Dhurjati, Lamb and Chester 1987). The limitations of each type must be respected otherwise their results can be misleading.

2.4 Fault Detection or Process Monitoring

Fault detection combines elements of multivariate modelling and prediction as well as diagnosing the cause (or causes) of changes or faults in a process. It utilizes knowledge about the process (either through some type of process model, statistical

characteristics or heuristics) for the monitoring and detection aspect, and the establishment of boundaries that act as decision rules to indicate that a malfunction or change has occurred and that diagnosis is needed. A further challenge is the updating of the process model or knowledge to adapt to acceptable changes in the process. Process monitoring methods are discussed in Himmelblau (1978) and Willsky (1976). Basseville (1988) provides a survey focusing on likelihood ratio approaches. Some general aspects of the topic are discussed below.

Regardless of the type of fault detection method used, the key issues used for assessment are: i) the types of failure modes that can be considered, ii) the complexity in implementation of the method, iii) its performance, as measured by the frequency of false alarms and delays in detection, and iv) the robustness in the presence of modelling error (Willsky 1976).

The trade-off is between complexity (e.g., expense as measured by implementation, cost of false alarms, etc.) versus performance and is obviously quite specific to the process being studied.

Problems in dealing with process data for fault diagnosis include: i) validating readings given by the measurement instruments, ii) compensation for time lags, iii) elimination of noise in instruments, iv) high interaction amongst process components making isolation of cause difficult in complex systems, and v) drifts in parameters (Himmelblau 1978).

Difficulties in early detection of dangerous states arise from: i) inability to monitor the entire "process state" as one is often limited to just temperature, pressure , flow and concentration readings, ii) the complexity and non-linearities of the process, and iii) the need to incorporate historical information about the process along with current measurements (King 1986).

As was the case for prediction models, fault detection methods form a spectrum stretching from all-theory mechanistic models to all-data statistical hypothesis testing. One example of a mechanistic model approach is the use of a kalman filter to obtain the minimum variance estimates of state variables of a mechanistic model consisting of ordinary differential equations. The state variables may describe system behavior that cannot be directly measured such as compositions in a reacting system. After comparing the available measured outputs with its model prediction, the Kalman filter uses the discrepancy between these two values to update its state estimates.

King (1986) gives an example of a direct method for fault detection using a Kalman filter where several "failure states" or filters are incorporated into the dynamic process model. When the estimates for these states deviate markedly from their normal values, a failure is detected. Since the Kalman filter predicts the location of the process at some future time, it allows for early detection of undesirable states.

The all-data statistical hypothesis testing approach brings us back to the control charts (Shewhart, CUSUM and EWMA) discussed earlier. Process variations are divided into two sources; random fluctuations (due to such phenomena as external environmental changes in temperature and pressure, internal mixing conditions or natural variations in raw materials), and non-random changes (which could be the result of operational moves, faulty instrumentation, off-specification raw materials, catalyst deterioration and so on). The statistical control approach assumes that the process will remain on-target "in a state of statistical control" unless a special event occurs. Hypothesis tests (in the form of limits or masks) are used to detect such events and this is followed by a search for an assignable cause (or causes) and correction of the process by removal of the cause or applying compensation for it.

In addition to the problems of colinearity amongst monitored variables and serial correlation (augmented by the presence of inertial elements, recycle and reactors in industrial processes) other difficulties have hampered the widespread use of control charts in process industries. Diagnosis and assignment of causes is very difficult except in the cases of gross abnormalities (e.g., equipment failures and improperly set control variables). It is also hard to assign causes to cyclical or level changes. Further, the goal of many process industries is not controlling an output to within a given range (as it is in parts manufacturing) but to maximize a variable (e.g., yield), thus the target values for the monitored variables can shift regularly as the point of optimum operation moves.

In the middle of this spectrum lie expert systems. These use both theoretical and empirical models in their fault detection and diagnosis schemes. Expert systems are often motivated by a desire to combine the diagnostic expertise of process experts with the high computational speeds of process computers to provide a quick and efficient way of diagnosing process problems. Not only are there cost and time savings associated with their use, but expert systems can also provide a fast transfer of expertise to new operating personnel and thus have potential as a training tool.

Expert systems require large amounts of knowledge coupled with an appropriate problem-solving method. Of course, these are dependent upon the types of faults one wishes to detect, the complexity of the process and the frequency at which assessments are to be made. The knowledge base is formed from continuity equations and physiochemical principles as well as heuristic knowledge or "rules" gleaned from process experts. Many different strategies have been developed for the problem-solving task and Shum et al. (1988) provides a list of recent work in this area.

Two expert system projects, the CATCRACKER and the FALCON, were developed specifically for fluidized catalytic cracking processes and are briefly discussed below.

The framework of the CATCRACKER expert system (Ramesh and Davis 1989) consists of three tasks; classification (deciding which of a number of fault hypotheses apply to the process when a change is detected), abductive assembly (finding a reason for an observed change based on abnormal deviations in variables known to create such a change), and data abstraction (abstracting high level information from raw data measurements). It employs a hierarchical structure of fault categories to speed up isolation of the problem and elimination of hypotheses with a possible 103 root causes forming the base. Detailed tests are performed and past data is used in the assessments as well as current data. Such complexity limits the speed at which diagnosis can be made. In fact, the CATCRACKER system is only capable of diagnosing slow response problems (time scales on the order of hours, shifts, or days).

The FALCON expert system (Dhurjati, Lamb and Chester 1987) starts with a set of 39 faults which are to be detected. The project was meant to demonstrate the utility of expert system technology in a commercial scale process and also to identify the resource requirements of the technology. Being smaller in scope and less complex in structure than the CATCRACKER system allows the FALCON system to make assessments every 15 seconds.

One very important issue which arose after testing the FALCON system with plant data was that

> . . . the number, placement and precision of the sensors on the plant determine to a large extent the magnitude of the fault that can be detected. A sensitivity analysis based on the methodology for fault detection can be used to specify limits on the maximum magnitude of faults that can be detected. Conversely, one can specify sensor placement and precision needed to diagnose faults of a prespecified magnitude. (Dhurjati, Lamb and Chester 1987)

Issues that have yet to be addressed with these two projects are validation of sensor data and the ability to handle interacting, multiple malfunctions.

Thus, the challenge in analyzing, modelling and monitoring large process data sets lies in going beyond the univariate approaches of control charts and traditional regression fitting but without having to resort to the sophisticated and time consuming approaches of theoretical models. PCA and PLS appear suitable for this challenge as they have demonstrated their ability to reduce large matrices of complicated data into low dimensional and easily understandable models despite being subjected to complications such as high colinearity amongst measurements (as many readings reflect the same underlying change in the process), presence of noise, and many more measurements than samples.

CHAPTER 3: THE FLUIDIZED CATALYTIC CRACKING PROCESS

This chapter introduces the fluidized catalytic cracking unit (FCCU). A description of the process is followed by an explanation of typical constraints and disturbances which define its operating window and the policies and control strategies which determine its operation. A brief literature survey of FCCU models is also included to illustrate what inadequacies still exist in this field.

3.1 Process Description

Fluidized catalytic cracking units are one of the most important units in petroleum refinery operations. They are used to convert the heaviest mid-third of the crude oil cuts into gasoline and diesel blends and other light products. A typical industrial FCCU consisting of a reactor and two-stage regenerator is shown in figure 3.1.

Atomized feed (after having been pre-heated by a fired furnace or exchanging with product slurry) is injected into a hot stream of regenerated catalyst at the base of the reactor riser. It immediately vaporizes and the cracking reactions take place as the catalyst-feed mixture travels up the riser (residence time typically being only a few seconds). The resulting degree of conversion and coke production are a complex function of feed quality and catalyst activity, as well as carbon level and metals content on the regenerated catalyst. In addition to thermal cracking of long hydrocarbon chains, many secondary reactions take place; isomerization of olefins, dehydrogenation of napthenes, and polymerization of aromatics which can stay on the catalyst and convert to coke. The reactions are endothermic, with most of the necessary thermal energy being provided by the hot regenerated catalyst.

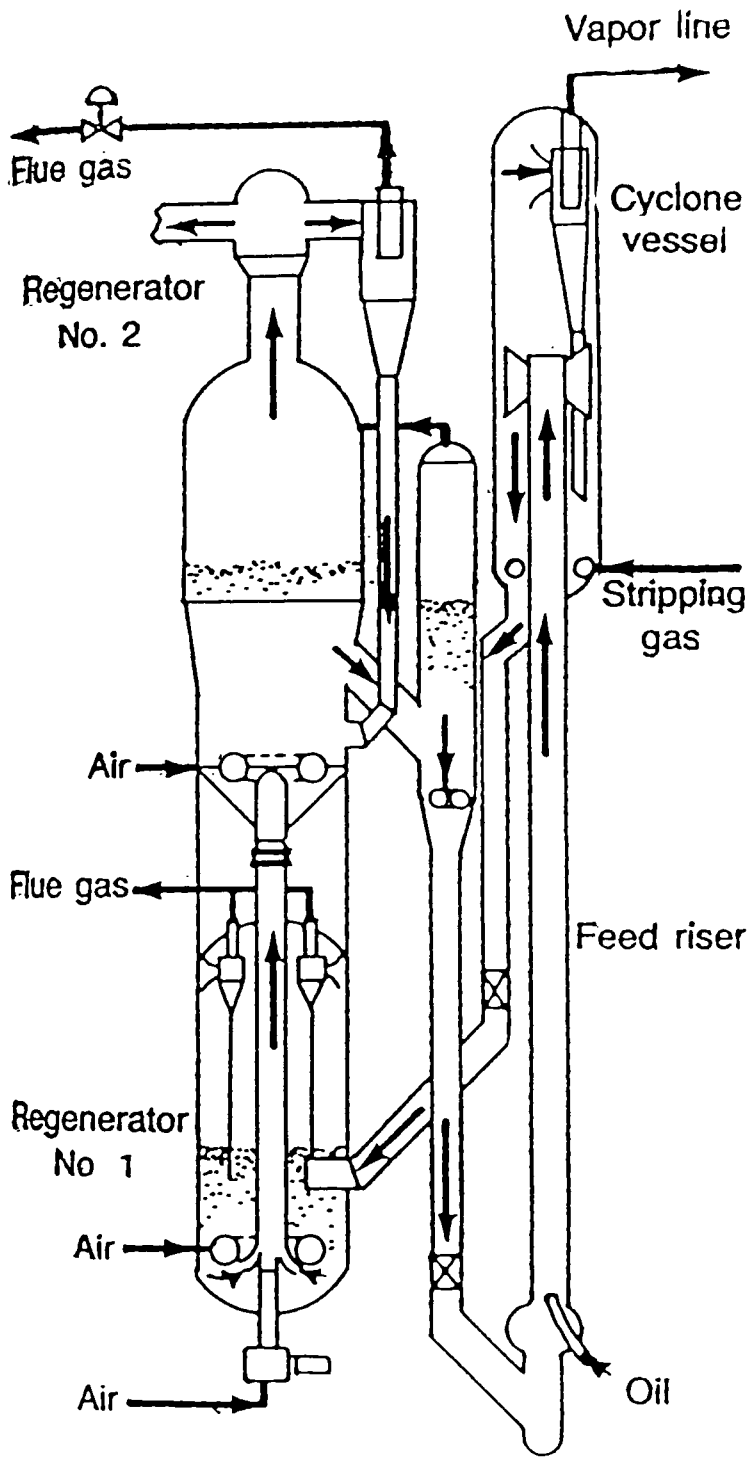


Figure 3.1: A FCC unit with a two-stage regenerator. (Dean, Mauleon and Letzsch 1982)

Above the riser in the reactor vessel, the product gases are separated from the catalyst by cyclones and sent to a fractionation section, as illustrated in figure 3.2. The spent catalyst is steam-stripped of any remaining hydrocarbons in the reactor bed before flowing to the regenerators where the coke is burned off and catalyst activity rejuvenated.

Although many FCC units only have a single regenerator, as depicted in figure 3.2, the unit studied in this work has a two-stage regeneration process, as shown in figure 3.1. The first stage operates at a relatively low temperature and is responsible for the burning off of hydrogen and removal of steam carried over from the stripping section, as well as partial conversion of the coke (C) on the catalyst (about 50%) to carbon monoxide and carbon dioxide (CO and CO_2, respectively). The complete conversion of C to CO_2 (known as "afterburning") is highly exothermic, releasing about three times as much thermal energy as the partial conversion of C to CO. The high temperatures accompanied by afterburning must be avoided in the first stage due to the presence of water (i.e., steam carried over from the reactor side); this can lead to hydrothermal degradation of the catalyst. Flue gases from the first stage regenerator are passed through a series of cyclones to remove any entrained catalyst and are then fed to a CO boiler where supplemental fuel and air are used to complete the CO oxidation and recover thermal energy via production of steam. The single regenerator of the FCCU in figure 3.2 is operated in the same manner as this first stage.

Catalyst from the first stage bed is transported to the second stage through a lift pipe with supplemental air. The second stage is constructed of high-strength high-temperature resistant alloys. This allows air to be supplied to ensure complete combustion of C and CO, and a thorough regeneration of the catalyst. Complete combustion in the second stage also prevents possible afterburning in downstream flue gas lines. Air supplied for combustion to both regenerators and for transport of catalyst from the first to the second stage is generally supplied by a single blower (compressor).

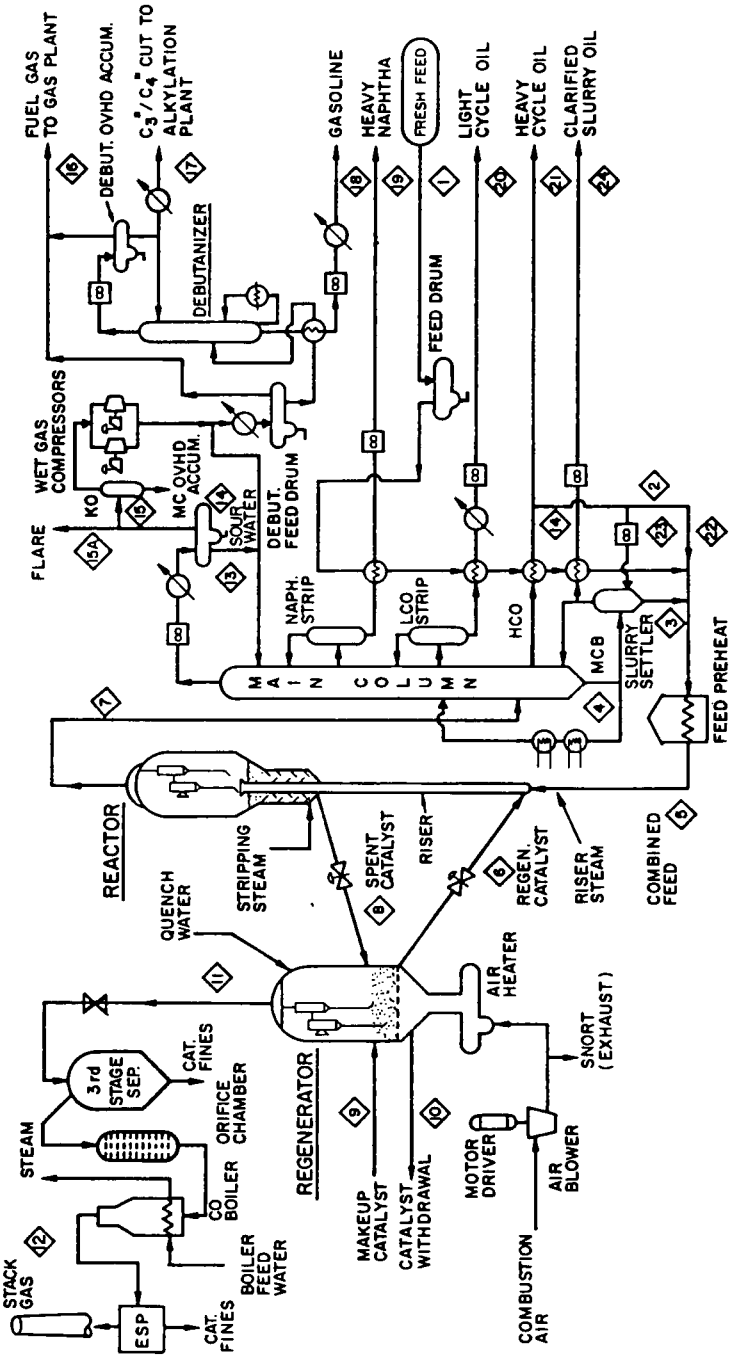


Figure 3.2: A typical FCCU showing fractionation section. (Venuto and Habib 1978)

The downstream fractionation section typically consists of a main fractionation column where the most heavy products (cycle and slurry oils) are separated. The fractionator overhead gas is compressed and then flows to a series of distillation columns which split it into fuel (dry) gas, propane, butane and gasoline product streams.

The complexity of the unit's operation arises from the heat, carbon and pressure balances which must be maintained to ensure safe operation. For stable reactor and regenerator temperatures, the quantity of heat generated by the exothermic combustion of coke in the regenerators must balance the amount required in the reactor to vaporize and crack the feed. This coupling of endothermic and exothermic reactions creates a high degree of dependence amongst the process variables.

To maintain the catalyst activity level, the carbon laid down during the cracking reactions must be burned off in the regenerators. Catalyst losses to flue gas are balanced by daily additions of fresh catalyst to the system. The pressure balance between the vessels ensures proper directional flow of the catalyst so that hydrogen-rich reactor gases do not combine with the air-rich regenerator gases. The pressure drops across the slide valves determine the actual catalyst flow rate.

3.2 FCCU Operating Window

Each FCCU has its own set of process constraints which define its operating window and is subject to numerous disturbances which can alter the ideal operating point in the window or the shape of the window itself. Discussion of theses issues and typical values for operating points can be found in Brice and Krikorian (1983) and Venuto and Habib (1978).

3.2.1 Constraints

Riser outlet temperature (ROT) is the key variable used to control conversion and thus has upper and lower constraints. If the temperature is too high, over-cracking causing increased light gas production takes place and could lead to flooding of the downstream fractionation unit. Too low a temperature generates little conversion and may cause fouling of the catalyst.

Feed quality plays an important role in defining the operating window. Heavy components or residues present in the feed require more energy to crack and lay down more coke on the catalyst. This can drive the ROT to its upper limit as well as pushing the unit to the mechanical constraints of its air blower(s) and fractionation unit. Arkun and Stephanopoulos (1980) show how the steady-state operating region of a (simulated) FCCU changes with feed quality; Palazoglu and Khambanonda (1987) do the same for both feed quality and changing feed rate.

Capacity of the air blower is limited by the maximum power that its motor drive can deliver. The power requirement at any time is influenced by the amount of coke to be burned in the regenerators, pressure in the regenerators and the ambient air temperature.

The wet gas compressor, located upstream of the product separation columns, is also limited by the power of its motor drive and affected by the production of light gases (or reactor pressure).

Maximum feed temperature is limited by the amount of heat which the pre-heater or exchangers can deliver, and by the minimum catalyst-to-oil ratio set by operations. If the feed is too hot, catalyst circulation must be reduced, which leads to poor conversion.

Bed temperatures in the two regenerators are limited by the metallurgy of their internals. The first regenerator limit is quite low, since it is operated to avoid afterburning. The second regenerator limit is much higher, since it is made of more heat-resistant alloys which allow afterburning to safely take place.

The concentration of carbon on the catalyst needs to be maintained at as low a level as possible to ensure adequate catalyst activity. This is influenced by the quality of the feed and is limited by the amount of air available for burning off the carbon.

3.2.2 Disturbances

The two most common and serious disturbances affecting FCCU operations are also the hardest ones to monitor and control; feed quality and catalyst activity.

Typical catalytic cracking feed comes from not one but several upstream processing units in a refinery. The feed stocks contain a mixture of hydrocarbon species that have a wide range of cracking rates. Several supply streams flow directly to the FCCU feed tanks and thus the composition of the FCCU feed varies with the operating policies or disturbances at these upstream units. The components of most concern are the heavy residues, known as "pitch", originating from the crude unit. Not only is pitch harder to crack, but it generally contains the most catalyst poisons of all the feed stocks. Since it is expensive and time-consuming to characterize the feed stock, only infrequent analyses are performed on it, leaving operations personnel to run the FCCU with only a rough guess of the feed quality.

Catalyst quality is also a crucial but complex and difficult variable to quantify. Catalyst activity is a function of the amount of carbon laid down during cracking and the amount removed in the regenerators, addition of fresh catalyst to the beds, and the degree of metals poisoning and hydrothermal degradation that may have taken place.

The effect of catalyst (separate from process or feed effects) on unit conversion and performance is evaluated by the micro-activity test (MAT). This measures the conversion obtained at standard conditions in a laboratory from a de-coked sample of catalyst. Changes in the MAT value can then be attributed to the catalyst alone and should also be reflected in the unit's conversion (providing all other variables remain constant).

Activity is not monitored on-line in FCC units and its impact on yield is not immediately evident (it may require several days to show any effect). Yet high activity is crucial for high conversion and the selectivity of the catalyst influences the product slate.

FCCU catalysts are particularly sensitive to metals poisoning. Deposits of nickel, vanadium, and to a lesser extent iron and copper, cause increased hydrogen and coke production. Sodium, lithium, calcium and potassium diminish the thermal stability of the catalyst and cause increased propane and butane production (Upson 1981).

Other disturbances include feed temperature (caused by a sudden change of feed source such as pulling from storage if an upstream unit reduces output), and ambient temperatures (which affect air blower capacities and cooling water temperatures).

3.3 Operational Policies

Unlike other industrial processes where production of a single product of consistent quality requires a single steady-state operating point, the FCCU is subject to a continually changing operating policy which means the FCCU is run in numerous operating regions. Two of the most typical operational changes are due to the seasonal shifts between gasoline and light cycle oil production, and unit feed rate changes, both of which are functions of market demand. For refineries which also produce asphalt on an irregular basis, the larger-than-normal swings in crude feed quality have an added impact

on FCCU feed quality and thus its operating point. Lee and Weekman (1976) show how the operating point of an FCCU jumps around its operating window, depending upon the constraints placed on the unit.

Hence, the FCCU must have a flexible and relatively large operating window to accommodate these operational demands. It makes defining a set of conditions which represent "normal" operations difficult as this is heavily dependent upon a number of conditions (market demands, feed stock, product slate, catalyst quality).

3.4 Control Objectives and Strategy

The control policy of an FCCU must be flexible enough to accommodate changing unit feed rates and disturbances in feed quality while maintaining relatively steady-state operation, so as not to upset other downstream units in the refinery. The most important objectives are to ensure safe operation (e.g., reverse flow protection on the catalyst circuit to prevent explosion of hydrocarbon-rich gases in the oxygen-rich regenerator), maintaining the heat and carbon balances within the reactor-regenerator circuit, maximizing conversion (subject to economics and unit constraints), overcoming the process variables' interaction to ensure smooth control, and to work as close as possible to unit constraints to maximize profitability (Brice and Krikorian 1983).

The control strategies in place determine which key variables move and why; this is important when trying to interpret the latent variables of PCA and PLS models.

The independent variables typically regulated are: riser outlet temperature, feed rate, feed pre-heat temperature, reactor pressure and catalyst activity (Venuto and Habib 1978). Changes in these variables reflect changes in operating policy or disturbances which are not controllable.

Controllable disturbances entering the process are signalled by fluctuations in the control loops' manipulated variables. In a conventional FCCU control strategy (Kurihara 1967) these are i) regenerated catalyst slide valve (controlling riser outlet temperature), ii) spent catalyst slide valve (controlling catalyst level in reactor), iii) wet gas compressor suction pressure (influencing reactor pressure), iv) second regenerator flue gas slide valve (controlling the pressure difference between reactor and regenerator) and v) air rate (controlling the percent excess oxygen in the flue gas).

Dependent variables (which respond to changes in the independent variables) include catalyst circulation rate, regenerator temperature, conversion and the air rate required to support combustion of coke deposits (Venuto and Habib 1978).

Unit operation is further complicated by a mixture of variables which respond very quickly to control and operational moves (e.g., catalyst circulation, heat balance, riser outlet and regenerator temperatures), and those which respond very slowly (e.g., catalyst activity, heavy oil quality which if recycled affects feed quality) (Brice and Krikorian 1983). Monge and Georgakis (1987) were able to show this through simulations of an FCCU.

Thus, the FCCU is a highly complex, highly coupled process which poses a great challenge for simple monitoring, modelling and analysis techniques. Its key disturbances are hard to measure and monitor. The most important process variables (gasoline octane number, catalyst activity, carbon on catalyst concentration) cannot be measured on-line but must be inferred. The unit's operating policy is continually changing with market forces, and constraints imposed by feed catalyst quality and mechanical limitations cause the process to move around its operating window. A mixture of fast and slow response modes add to the complication in defining stable, steady-state normal

operating conditions. Although these issues may complicate the analyses, they are typical of the types of problems posed by industrial data sets and PCA and PLS should be better at accommodating them than traditional multivariate techniques.

3.5 FCCU Models

In this section, complete FCCU models will be discussed to illustrate what is currently available in the literature and what inadequacies still exist in this area. Kinetic models of catalytic cracking will not be covered since the literature is too expansive and detailed in this area to be reviewed here.

Modelling a process like an FCCU requires trade-offs between complexity and reliability, fundamental process knowledge and empiricism, and involves problems in scaling from pilot to commercial scale units. A high degree of sophistication and detail requires a high level of expertise (at an equally great expense) while too simple a model may not be reliable or serve its purpose well. Although every model contains some degree of empiricism, the less there is of this, the greater will be the generality of the model. This reduction comes about through increased understanding of the process fundamentals. It is also generally difficult to get reliable quantitative information on commercial units. Since process performance is equipment sensitive, scaling up pilot plant data to match commercial scale units can be difficult (McDonald and Harkins 1987).

Elnashaie and Elshishini (1990) recently reviewed the following eight FCCU models developed over a span of twenty-seven years; Luyben and Lamb (1963), Kurihara (1967), Iscol (1970), Lee and Kugelman (1973), Elnashaie and El-Hennawi (1979), Farag and Tsai (1987), Edwards and Kim (1988) and Elshishini and Elnashaie (1990).

The first four models are variations on CSTR representations of the reactor and regenerator. The Luyben and Lamb, and Kurihara models involve only three component species (typically gas oil, gasoline and other products and coke) while Iscol and Lee and Kugelman do not use any reaction network. Elnashaie and El-Hennawi take into account the two-phase nature of the reactor and regenerator by modelling the bubble phase and dense phase of each separately. They also use a three component reaction model. Farag and Tsai developed a correlation model which predicts the trends of operating variables on reactor product yields (fuel gas, propane, butane, gasoline, light gas oil and coke). Their paper also notes other empirical models. The Edwards and Kim model is a proprietary one. Elshishini and Elnashaie's model is a modification of the Elnashaie and El-Hennawi one and includes empirical calculations for light hydrocarbons (i.e., separates light gases from the single coke and gas term) and heavy cycle oil recycled as feed. Some additional models not covered by the review are discussed below.

McGreavy and Smith (1984) use a three lump model for reaction in their quasi-steady state riser. It includes a model for the stripper (a lumped capacitance model) and catalyst circulation. The regenerator is modelled as a two-phase fluidized bed and includes afterburning, catalyst entrainment and recirculation effects. Lee and Groves (1985) combine the adiabatic plug flow reactor model of Shah et al. (1977) and perfectly mixed tank regenerator model of Errazu, de Lasa and Sarti (1979) along with a three lump model for the reaction network.

The riser portion of the model by McFarlane et al. (1990) allows for varying feed quality (i.e., tendency to produce coke) but only predicts two products; yield of wet gas (hexane and lighter) and the amount of coke deposited on the catalyst. The regenerator is a two-phase model, and simple models for lift air and combustion air blowers, catalyst circulation, main fractionator and wet gas compressor are included.

Many of the above works using reaction models define their products so widely that only limited information is available on how process variables affect gasoline yield and the other six to seven products of a commercial FCCU. Kraemer, Sedran and de Lasa (1990) have developed an eight lump model (dividing the products up as: gasoline, butane and lighter plus coke, and light ($220°C - 343°C$) and heavy ($343°C +$) cuts of each of paraffins, napthenes and aromatics), but it has yet to be incorporated into a published FCCU model.

Only some models allow study of the effects of feed and catalyst quality on process operations and products and since many models are fitted to specific industrial units, they are only valid in narrow operating ranges and only for their respective unit. No model for a two-stage regenerator has yet been published in the literature.

CHAPTER 4: PCA, PLS AND HIERARCHICAL PLS - DESCRIPTIONS AND ISSUES

In this chapter the methods of PCA, PLS and hierarchical PLS are briefly outlined along with recent examples of their implementation in analysis, modelling and monitoring roles. This is followed by a discussion on appropriate model dimensionality, data pretreatment issues which affect the outcome of these analyses (scaling, dynamic data, time delays, drifts in the data set, normalization, reference set selection, outliers, and hazards of modelling with historical data) and a set of tools available for analyzing the results.

4.1 Introduction

PCA and PLS are from the family of pattern recognition techniques which build distribution-free empirical models where causality is based on correlation (Geladi 1988). The goal is to create a low dimensional representation of a high dimensional matrix (or matrices) by finding a small number of latent variables (LVs), linear combinations of the measured variables, which correspond to the directions of greatest variation in the data set.

Latent variables can be thought of as new "meters" or measurement instruments which monitor the dominant variations or changes taking place in the process. They may be more directly related to physical or chemical phenomena than single process variables collected. Attractive features of LV models are the orthogonality in the LVs (which allows for easy interpretation, display and subsequent analysis of the results) and eigenvector

properties which assure computational convergence (Wold et al. 1987b). Latent variables can be plotted against each other to form two-dimensional planes or "windows" which help to visualize the relationships amongst the variables and samples.

The strength of PCA, PLS and hierarchical PLS lie in their capacity to deal with highly colinear data sets, cases where the number of measurements made is greater than the number of samples available, and sets where data is missing or may contain outliers. The NIPALS algorithm (Wold et al. 1984) was used for all three methods. It is the basis of the software SIMCA and was run from a MATLAB 386 version written by Skagerberg (1990). Complete details on all the calculations can be found in Wold et al. (1984).

The following sections briefly discuss the mathematics of each method, and highlight recent applications.

4.2 Principal Component Analysis

PCA, (also known as singular value decomposition, eigenvector analysis, characteristic root analysis, characteristic root analysis, latent vector analysis and Karhunen-Leuwe transformation) works with a single data block and yields latent variables which purely summarize the variance information amongst the variables. Figure 4.1 shows a latent variable fitted to an X block consisting of three variables and n samples. The latent variable lies in the direction of greatest variation in the three dimension space that the variables span.

Mathematically, PCA decomposes the data matrix X into a set of vectors p and t called loadings and scores:

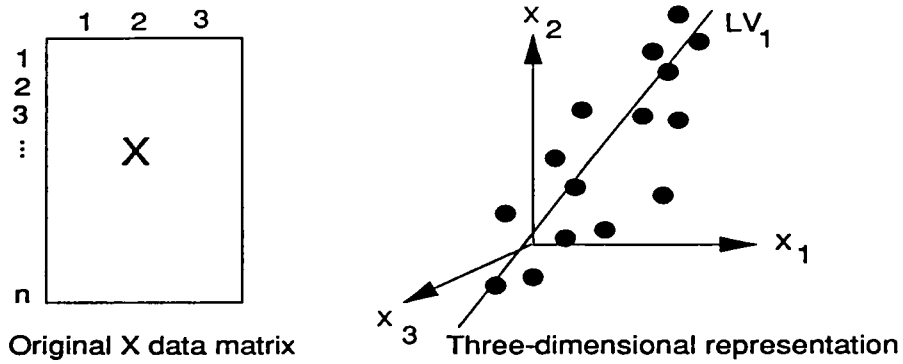


Figure 4.1: Visual representation of fitting a latent variable (LV) to a data block (containing three variables).

$$X = t_1 p_1^T + t_2 p_2^T + \ldots t_A p_A^T + E_A \qquad (4.i)$$

t_a = vector of scores (or values) for each sample for dimension a

p_a = vector of loadings (or weights) of each process variable for dimension a

E_a = residual matrix after extraction of "a" latent variables

or principal components

A = number of latent variables in the model

The contribution of each process variable to the latent variable is represented through the loadings in vector p. Each sample has a score, t, associated with it which simply represents the value of the latent variable for that set of measurements.

The NIPALS algorithm (illustrated in figure 4.2) starts by selecting the column with the largest variance in X as an initial guess of t, then using the X data, calculates p:

$$p^T = \frac{t^T X}{t^T t} \qquad (4.1)$$

(normalize p to unit length)

A new t vector is then calculated from p and X and compared to the old one.

$$t = \frac{Xp}{p^T p} \tag{4.2}$$

If t has not converged, one returns to equation 4.1 and repeats the calculations. Once convergence is reached, the original data matrix can be described in terms of the score vector t, loading vector p and residual block E as shown in figure 4.3:

$$X = tp^T + E \tag{4.3}$$

The PCA latent variables are also known as "principal components" (PCs). The first PC describes the direction of greatest variability in X. NIPALS then makes E the new data matrix (since it represents all the information left unmodelled) and repeats the calculations for the next dimension.

Dimensionality reduction is achieved by selecting the value of A in equation 4.i to be much smaller than the number of variables present in X and such that no significant process information remains in the matrix E_A (i.e., E_A only represents random noise).

For data scaled to unit variance (discussed in section 4.6.1) the PC analysis is essentially finding the eigenvectors of the correlation matrix (Chatfield and Collins 1980). Wold, Esbensen and Geladi (1987c) also show how PCA is directly related to singular value decomposition.

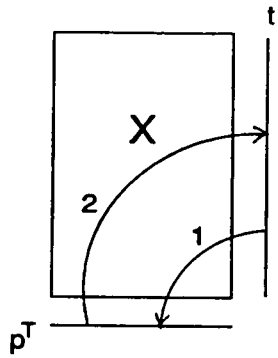


Figure 4.2: One dimension of the NIPALS PCA algorithm. The numbers represent the order of the calculations.

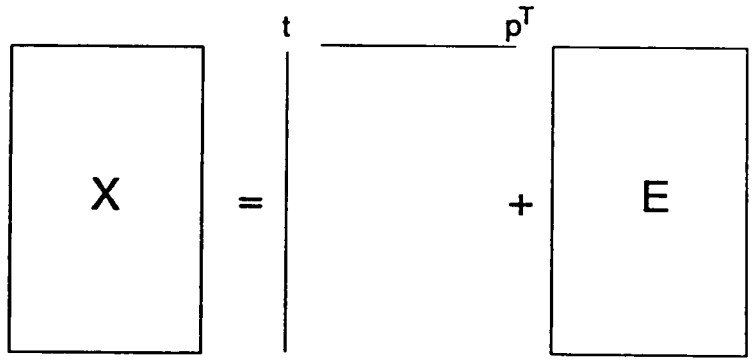


Figure 4.3: Vector representation of the X data block after calculation of one latent variable.

The analysis is not an end in itself, but a way to reduce the dimensionality of a problem before carrying out further analyses. Thus, the objectives of PCA encompass simplification, data reduction, outlier detection, variable selection, classification, prediction and modelling (Wold, Esbensen and Geladi 1987c).

Wold et al. (1984) give several examples and references to work using PCA for such purposes. Typical applications in the field of chemistry include classification of compounds according to their type of biological activity, chemical reactivity and structure. Non-chemical applications include classification of ecological systems, economical systems and accounting tables.

Modelling and analysis of data sets collected from industrial processes has thus far been limited. Wise et al. (1988) used PCA to model a monitoring space for a liquid fed ceramic melter. The system was designed to warn operators of process upsets, off-specification operating conditions and to help in the identification of failing sensors. Piovoso, Kosanovich and Yuk (1991) built a similar monitoring space for the solution area of a polymer yarn fabrication process, with much the same objectives in mind.

4.3 Partial Least Squares or Projection to Latent Structures

When one has two sets of variables (X and Y) and wishes to predict one set of variables (Y) from the other (X), PLS is an approach to use. Latent variables for each block are selected simultaneously and rotated so that they compromise between explaining the variance in X and predicting Y.

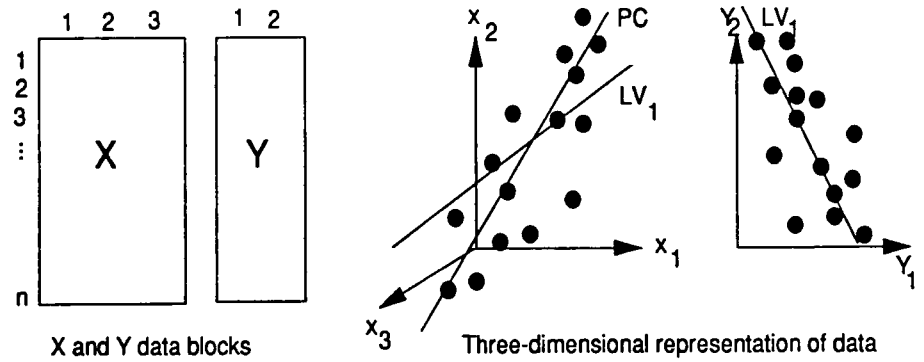


Figure 4.4: Fitting a LV to an X and Y data block using PLS.

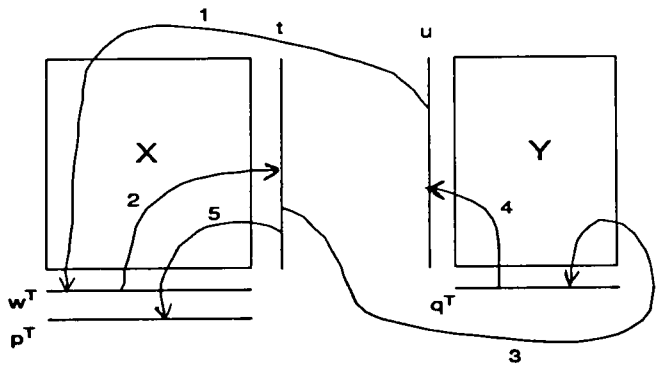


Figure 4.5: One dimension of the NIPALS PLS algorithm. The numbers represent the order of the calculations.

Figure 4.4 illustrates this for a data set composed of three x variables and two y variables. The principal component in the X space (labelled as PC) represents the vector which describes a maximum amount of variance, however, it must be rotated to maximize its correlation power with the Y space resulting in the LV_1 fit shown.

The NIPALS PLS algorithm is outlined in figure 4.5 and the reader is directed to the paper by Wold et al. (1984) for complete details of the calculations. Briefly, the column of maximum variance in the Y block is selected as an initial guess for the vector u, and vectors w^T, and q^T are then calculated in series followed by an update of u:

$$w^T = \frac{u^T X}{u^T u} \tag{4.4}$$

(normalize w to unit length)

$$t = \frac{Xw}{w^T w} \tag{4.5}$$

$$q^T = \frac{t^T Y}{t^T t} \tag{4.6}$$

(normalize q to unit length)

$$u = \frac{Yq}{q^T q} \tag{4.7}$$

If t has not converged, one returns to equation 4.4 and repeats the sequence. Once convergence of t has been reached, the X loadings p^T and regression coefficients b are determined:

$$p^T = \frac{t^T X}{t^T t} \tag{4.8}$$

$$b = \frac{u^T t}{t^T t} \tag{4.9}$$

The X and Y blocks can now be predicted from:

$$X = tp^T + E \tag{4.10}$$

$$Y = btq^T + F \tag{4.11}$$

where

$$b = \frac{u^T t}{t^T t} \tag{4.12}$$

= biased regression coefficients for one dimension

(if q is not normalized, b equals unity).

For a given model dimensionality, overall biased regression coefficients, B, can be calculated as follows (Wold et al. 1987a):

$$B = \frac{W}{P^T W} \tag{4.13}$$

where

W = block containing all w vectors

P = block containing all p vectors

The residuals left over from fitting these models, E and F, are used as the X and Y spaces for the next LV extraction.

A very important feature of the results is that the t and w vectors are orthogonal. This allows one to form planes or windows which illustrate the reduced space as modelled by PLS.

Application of the PLS method first started in the fields of organic and analytical chemistry where multivariate data was abundant but working physical models were scarce. Its main role involved calibrating output from data-rich analysis techniques such as NIR, spectroscopy and chromatography to the properties or responses of samples being studied. PLS soon spread to the related fields of biology, clinical chemistry, medicine, food research, biotechnology, quantitative-structure activity relationships, and

pharmacology. Today, it is being applied in such diverse fields as education, psychology, management science, economics, political science and environmental science (Geladi 1988).

Kresta, Marlin and MacGregor (1991b) used PLS to develop an inferential control scheme for a simulated multicomponent distillation column. PLS has also been combined with PCA to build an industrial multivariate monitoring space. In the yarn fabrication example (mentioned under PCA) PLS was first applied to the data set to remove the known dominant effects of feed rate on process variability. PCA was then applied to the residuals of the PLS model to develop the "fingerprint" or portrait of the normal operating region (Piovoso, Kosanovich and Yuk 1991). However, as with PCA, little work involving PLS analysis of industrial data has been reported.

4.4 Hierarchical PLS

Hierarchical PLS, as the name suggests, is a modification of the PLS method where the number of blocks used for either the X or Y space (or both) is greater than one.

The original algorithm developed by Wold et al. (1987b) can be thought of as an expansion of PLS where single measured variables in the X and Y spaces are replaced by blocks of measured variables. In the case of "a" X blocks and "b" Y blocks, as shown in figure 4.6, for each dimension the algorithm iterates on finding the score and loading vectors most descriptive of each block, collecting the score vectors t_a and u_b from X and Y spaces into two composite matrices T and U, and performing a NIPALS-PLS round on these two blocks to update the consensus vectors t and u:

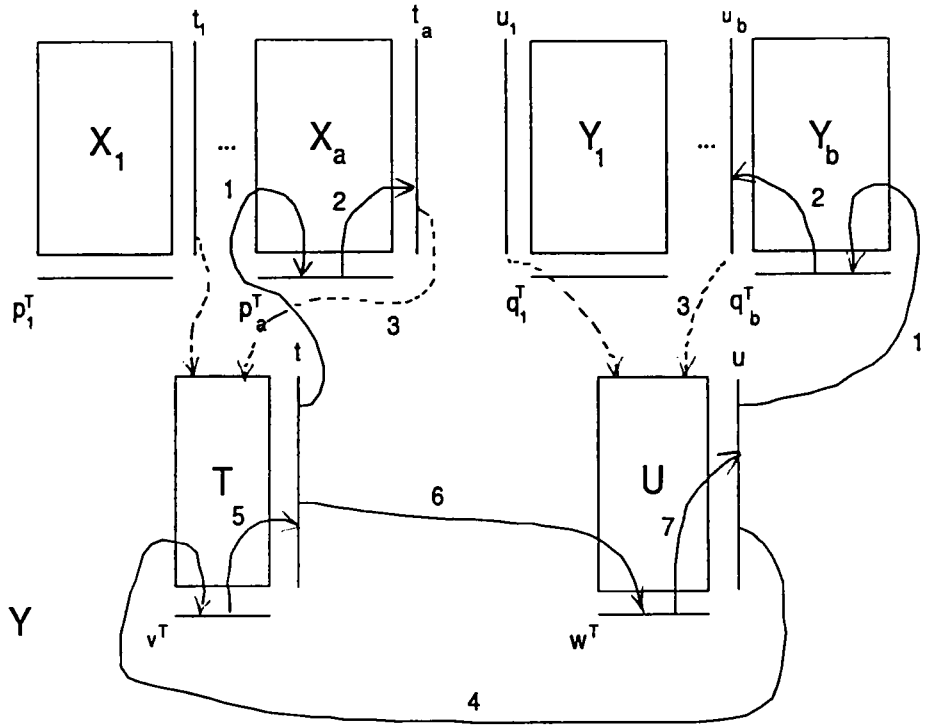


Figure 4.6: One dimension of the NIPALS hierarchical PLS algorithm. The numbers represent the order of the calculations.

Initial guesses of the consensus t and u vectors are made, for instance, by selecting the column with the largest variance in any X_a matrix for t and any Y_b matrix for u.

For each X block, the following vectors are calculated:

$$p_a = \frac{X_a^T t}{t^T t} \tag{4.14}$$

$$t_a = \frac{X_a p_a}{m_a} \tag{4.15}$$

where

m_a = number of columns in block a

The t_a vectors are collected into the consensus matrix T.

For each Y block:

$$q_b = \frac{Y_b^T u}{u^T u} \tag{4.16}$$

$$u_b = \frac{Y_b q_b}{m_b} \tag{4.17}$$

where

m_b = number of columns in block b

The u_b vectors are collected into the consensus matrix U.

A PLS round with T as X and U as Y is made to update the consensus vectors t and u, and calculate the block weight vectors v and w:

$$v = \frac{T^T u}{u^T u} \tag{4.18}$$

(normalize v to unit length)

$$t = \frac{Tv}{v^T v} \tag{4.19}$$

$$w = \frac{U^T t}{t^T t} \tag{4.20}$$

(normalize w to unit length)

$$u = \frac{Uw}{w^T w} \tag{4.21}$$

Convergence is checked on all t_a vectors as well as the consensus vector t. If this has not been achieved, the steps are repeated from equation 4.14.

One should note that the individual t_a vectors are not orthogonal and cannot be used to form monitoring planes.

The final consensus vector t plays an important role not only in the prediction of Y but also in the modelling of all the X blocks. The predictions for individual blocks are calculated, as follows:

$$X_a = t p_a^T + E_a \tag{4.22}$$

$$Y_b = t q_b^T + F_b \tag{4.23}$$

A modified approach of the above algorithm by Wangen and Kowalski (1988) allows for more complicated relationships amongst the blocks. Their approach is discused in greater detail in section 8.3.

As yet, no applications of the hierarchical PLS algorithm have been published.

4.5 Determining Model Dimensionality

An important issue in model building is determining the number of latent variables or components to use, as this has a significant effect on a model's resolution and predictive power. If the underlying relationship between the X and Y spaces in PLS (or within the X space in the case of PCA) is a linear one, the number of LVs needed to describe this should represent the number of independent phenomena taking place within the data set: non-linear phenomena will require extra LVs to describe their non-linearities (Geladi and Kowalski 1986).

The maximum possible dimensionality of a model is equal to the maximum number of columns or rows in X. However, since most process data sets contain colinearities, PCA and PLS are able to describe the significant variations in a data set using fewer latent variables.

If the precision of the measurements are known, one could calculate as many LVs as needed to reduce the matrix standard deviation to correspond with the precision standard deviation (Wold et al. 1984). However, this presumes that all the systematic chemical and physical phenomena in the data matrices can be modeled by a set of latent variables, and such an assumption can be quite questionable (Wold, Esbensen and Geladi 1987c). If there is unknown model error involved as well, then the above approach might lead to overfitting. Clearly, some other (and more rigorous) criteria is needed, such as cross-validation.

4.5.1 Cross-Validation

Cross-validation was developed to answer the question of what is an appropriate rank for a PCA or PLS model. It tries to estimate how much of the data is "signal" and how much is "noise" and to maximize the predictive power of the model (Wold 1978). For each LV calculated in an analysis, a random portion of the samples (say, $\frac{1}{4}$) is excluded from the X block along with their corresponding rows from the Y block; the LV is calculated from the remaining samples. Using the excluded X data and the LV, values for the excluded Ys are predicted and the predicted sum of squares, $PRESS_i$, is calculated:

$$PRESS_i = \sum_{n}\sum_{m}(\hat{Y} - Y)^2 \tag{4.24}$$

where

n = number of samples (rows) in the Y matrix

m = number of variables (columns) in the Y matrix

These steps are repeated i times, leaving out a different portion of the data each time until all data has been omitted once. The values of the $PRESS_i$ from all i repetitions are summed to yield the total PRESS for this LV. This value is then compared to the predicted value of the Y space, $\hat{Y}_{nm}$, using all previous latent variables:

$$SS = \sum_{n}\sum_{m}(Y_{nm} - \hat{Y}_{nm})^2 \tag{4.25}$$

where

Y_{nm} = measured Y value

$\hat{Y}_{nm}$ = predicted Y value using previous LVs

An intuitive criterion is applied to the following ratio:

$$CSV/SD = \sqrt{\left(\frac{PRESS}{SS}\right)} \tag{4.26}$$

If the value of CSV/SD is less than 1.0, the LV has significantly reduced the sum of squares (SS) in the Y space to warrant its use in the model. The paper by Wold (1978) describes the cross-validation process in detail.

Cross-validation is slightly conservative (i.e., it often leads to too few dimensions in terms of statistical significance) which is a positive attribute because the data are not over-fitted (Wold, Esbensen and Geladi 1987c).

4.6 Major Issues

Due to the nature of the soft modelling methods and their application in this work to a historical process data set, there are several key issues which need to be considered before carrying out the analyses. These are discussed in the following subsections.

4.6.1 Scaling

Since the LV models are fitted to the data using the criterion of least-squares, variables with a large variance will dominate the models. When process measurements span several orders of magnitude, large changes in relatively unimportant variables (e.g., cooling water flow rate) can swamp out smaller yet more significant changes in important variables (e.g., riser outlet temperature). Hence, some type of scaling of the data is necessary.

Since scaling directly affects a variable's relative variance, it also indirectly influences the nature of the LVs, resulting in the LVs not being a unique characteristic of the data (Chatfield and Collins 1980). There are numerous ways in which scaling can be done, such as i) scaling to give each variable equal variance (auto-scaling), ii) scaling by importance, or iii) scaling by product quality or controller ranges, to name a few.

Auto-scaling is typically used when one does not have a clear idea which variables are most important. It essentially scales each variable to equal importance (or influence) in the model by subtracting the column mean from each variable and then dividing by the standard deviation of the column. This also moves the coordinate system of the analysis to the center of the data (Wold, Esbensen and Geladi 1987c).

The main concern with this approach is that nearly constant variables with small variance will be scaled up in importance. However, if their variance is due mostly to noise, these variables should not contribute heavily to the model, only add to the variability which must be explained in the X (or Y) matrix.

Scaling variables by importance can reduce the risk of having meaningless, noisy variables dominate a model. However, the investigator must have a good idea of which variables are or may be important and this destroys some of the beauty of LV analyses, namely, extraction of information from the data without first imposing a

structure (in this case, through scaling) upon it. Scaling by product quality or controller ranges requires a more in-depth knowledge about the process and instrumentation in place but should yield a model which is easier to interpret.

This idea could also be approached from another angle, that is, to scale down variables whose variance is known to be mostly noise. A rule of thumb suggested by Wold, Esbensen and Geladi (1987c) is that

> . . . if the standard deviation of a variable over the data set is smaller than about four times its error of measurement, leave that variable unscaled. The other variables may still be variance-scaled if so desired.

In some cases, it may make more engineering sense to scale blocks of variables in the same way in order to retain the information about the relationship amongst them. The temperature profile of a distillation column is an excellent example (see Kresta, MacGregor and Marlin 1991a). Or, to give each variable in a block of similar type the same total variance, divide each variable by its standard deviation times the square root of the number of variables of that type in the block (Wold, Esbensen and Geladi 1987c).

Clearly, the scaling method used should be consistent with the aims of the data analysis and each case will have to be assessed individually for the most suitable approach.

4.6.2 Effects of Dynamics and Time Delay in the Data Set

Since dynamic data may introduce non-linearities into a data set (particularly in the case of multivariate systems) a latent variable model will need more dimensions to model this behaviour than the true dimensionality of the problem. Performing auto-correlations on the data can give the investigator a feel for the magnitude of the time constants for each variable if lags exist. However, the absence of significant auto-correlation lags does not ensure the data is steady-state, only that one has

snap-shot data. If the dynamics amongst variables are significantly different enough, they could also disrupt the correlation structure being modelled and this can have a serious effect on the model resolution and predictive power.

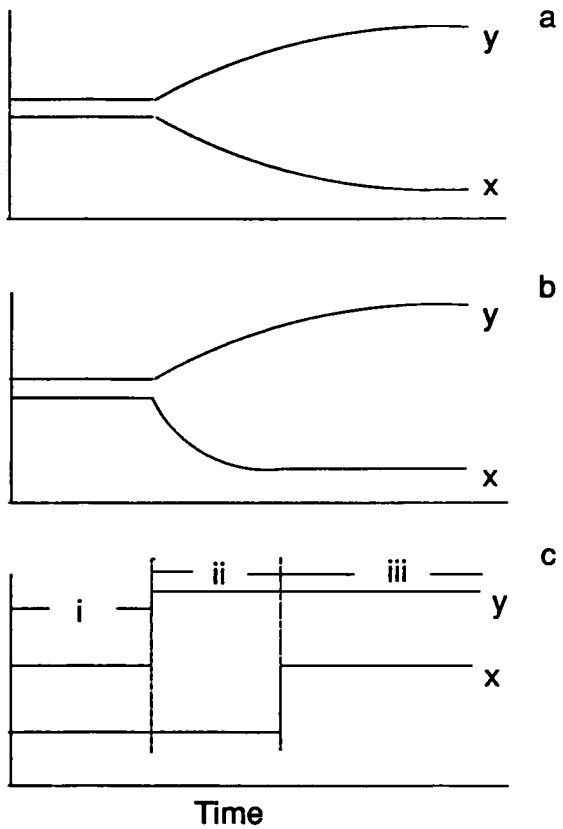


Figure 4.7: Effects of dynamics and time delay on the relationship between an x and y variable: a) same dynamics for both the x and y variable - relationship remains the same throughout dynamic period, b) different dynamics - relationship between x and y constantly changing, c) effect of time delay - (i) and (iii) show same relationship, while (iii) contains a different relationship.

Two possible dynamic situations are illustrated in figure 4.7a and b. If the y variable and a correlated x variable have the same dynamics (figure 4.74a) such that the relationship between the two variables remains the same throughout the dynamic period, then a LV model built from either steady-state or dynamic data should give the same result. The presence of the dynamic data does not have a detrimental effect on the model.

However, if the two variables have different dynamics, such as different time constants or response curves (as shown in figure 4.7b) such that the relationship to be modelled is constantly changing throughout the time series data, this leads to a complex model. A large number of components will be required to fit it and it may also be difficult to interpret the meaning of individual LVs. If the dynamics are very different over a given period, and the length of this period is significant relative to the amount of steady-state data present, then the presence of such dynamics in the data set can have serious effects on the resultant model.

A third effect to be considered is that of time delay. Figure 4.7c shows a fictitious time series plot of an x and y variable correlated in time but exhibiting a delay. For the time periods (i) and (iii), the data exhibit the same relationship, but in period (ii) a new relationship is introduced. This "transition" period can have serious effects on the steady-state model if the delay is large relative to the lengths of the steady-state periods. Cross-correlation calculations between the x and y variables can be used to check for such relationships, and if found, parts of the data can be shifted to align it and eliminate the time delays. Data can also be replicated then shifted so that there are no gaps in the time history, although this increases substantially the amount of data to be processed, and may not be a viable alternative for all analyses.

4.6.3 Drifts in the Data Set

Another important consideration is the presence of drifts in the data set. When data is collected in a class-wise fashion and measuring instruments or other sampling conditions slowly change over each class period, the samples will produce a class difference which is, in part, due to the drift and attempts to separate the contribution of the drift from the class differences will be difficult (Wold et al. 1984).

Ideally, the data collection should be designed, thus avoiding confounding with time trends, but when working with straight process data, one does not have this luxury. One possible way of reducing this problem would be to ensure that each class or event in the reference set contains samples from both the beginning and end of the set, although this is may not always be possible.

4.6.4 Normalization

Often the investigator knows that some mathematical relationship should hold for a set of measurements (such as a material balance), but due to measurement error or losses, the information does not fit exactly. One might then be tempted to normalize (weight) the data to make it fit, or to calculate a missing value from the others to fit the desired relationship. This forces a correlation upon the data and may mask information that would otherwise give some insight to the process.

Such procedures should be avoided as independent measurements of all variables lead to better parameter estimates and also provide information about the suitability of the model being entertained. This issue is discussed further by Box et al. (1973), Holly, Cook and Crowe (1989), and Wold et al. (1984). No normalization calculations are performed on the industrial data studied in this thesis.

4.6.5 Reference Set Selection

A set of data used to build a classification model is called the reference or training set. It is important that the reference set be selected such that the characteristics or classes one wishes to model are clearly defined and that one can get a pattern into which future values can be fitted (Wold et al. 1983).

Some necessary assumptions which apply to the reference set in order to achieve reasonable results are as follows:

* there are no strong outliers or subgroups present
* data homogeneity exists within a class (i.e., samples within a particular class must be similar in some way)
* the variables are monotonically related to the degree of similarity (i.e., the majority of the data measured on the samples must in some way be related to this similarity) (Wold et al. 1983)
* data should not be highly non-linear (variables can be transformed before performing the analysis to achieve a more linear data matrix (Geladi 1988; Martens, Wold and Martens 1983)
* when using process data, the data must properly span the operating space

To develop a model suitable for monitoring the operating behaviour of a process, as in multivariate SPC applications, the reference set must represent all those events or changes in the process which one considers normal. If the range of acceptable operations is too narrowly defined, the abnormality flag will be triggered too frequently and perhaps without due cause. If the range is too broadly defined by the data, little information will be provided about the process state and only gross changes from the process "norm" will be flagged.

Each class should contain at least five representative reference samples, and preferably up to ten or twenty (Wold et al. 1984). Keep in mind, however, that the idea of "classes" is a man-made construct, a type of structure imposed on the data, whereas the samples have a real existence. If it is found later that the resolution of the monitoring space is poor, one should consider whether or not the proposed classification (e.g., normal versus abnormal) is a good one or not in addition to questioning the information content of the measurements selected (Wold et al. 1984).

For developing prediction models, the reference set data should be collected according to a statistical design. Industrial processes, however, pose many problems for this strategy. The multivariate nature of processes such as the FCCU make it difficult to change only one variable at a time while holding the others constant. A factor believed to influence the output may not be easily isolated and changed (such as catalyst quality which can have a long time constant and is heavily influenced by feed quality), and the variables available to the investigator for manipulation in a designed experiment may not be those which are believed to influence the system (Wold et al. 1986).

If statistically designed data collection is not possible, then, at the very least, the reference set must span the process region expected to be found in the test data. The variance (and hence the error) of predictions arising from passive models grows as one moves further away from the centre of the reference set (Draper and Smith 1981). Thus, extrapolation can be expected to yield poor results.

4.6.6 Outliers

Sometimes a reference set can contain a sample which is an outlier; it does not truly represent any of the classes or events being modeled. Determining whether or not an observation is an outlier is very tricky. One does not want to automatically discard any singular point which appears to deviate from some preconceived trend because the

deviation may actually contain valuable information about the process. The deviation may also suggest that the preconceived model is wrong, a situation which is quite possible (Himmelblau 1970) particularly with undesigned data sets.

Due to the least-squares property of PCA and PLS, outliers severely influence the model by pulling it in such a way as to fit them closely (Wold et al. 1984). Thus, checking the residuals of fitted samples (e.g., performing an F-test on a sample's standard deviation of error compared to the total standard deviation) may not reveal the true outliers. In fact, they may have a smaller residual than samples which truly belong in the reference set (Wold et al. 1984).

Fortunately, such outliers can usually be detected in the T score plots. One can then use knowledge about what happened with the process during the sample periods to verify whether or not the sample is a legitimate part of the reference set.

The Mahalanobis distance is another tool which has been used (Piovoso, Kosanovich, and Yuk 1991) to determine whether a sample is a member of a reference class. It assumes the class population has a multivariate normal distribution, which thus imposes an ellipsoid shape onto the reference model (with the population mean sitting at its centroid). The size of the ellipse is defined by a chi-squared value for a user-specified confidence level and appropriate degrees of freedom and the distance measured is between its centroid and the sample location in the reduced subspace (Nilesh and Gemperline 1989). Due to the assumptions which must be made about the data, the Mahalanobis distance is not used in this work.

If the number of latent variables used in the model is greater than three, Martens (1985) suggests calculating a sample's leverage. Leverage indicates the degree of dominance of a sample in a model (but not whether that dominance is good or bad).

Once again, the retention of a "hypothetical" outlier depends upon the model usage and each sample which has a high leverage should be evaluated on this basis rather than being discarded automatically.

4.6.7 Hazards of Modeling with Historical Data

Models built from historical data are valid only within the range of the data used in the reference set. Extrapolation of a passive model is extremely dangerous due to the increase in prediction error that occurs as one moves away from the centre of the reference data (Draper and Smith 1981). Such models are also only valid as long as the process remains physically the same and the operational strategies do not change. Thus, the data set used for modeling must reflect the same process situation(s) which will exist when the model is used.

Complications can arise from the presence of inconsistencies in the data set over time, the presence of control loops, semi-confounding of effects, and attempts to draw causal relationships from the regression models. These issues apply to any regression equations fitted to undesigned data sets (Box, Hunter and Hunter 1978) and are briefly discussed below.

4.6.7.1 Inconsistent Data

A long historical data set can contain many inconsistencies due to small changes over time in instrument errors, calibration drifts, operator changes and so on. Although some of these changes can be accounted for (and are recorded in operating records) many are not. By introducing inconsistent structures into the data set, such samples can greatly dilute the resolution and power of the model.

4.6.7.2 Presence of Control Loops

Controlled variables pose an interesting problem for models built by regression methods. The variance of their movement is typically limited because these variables often have a significant effect on the process. With limited range, however, they may appear to have no effect on the output space and thus will typically not have large coefficients in the regression model. One has essentially modeled not the relationship between the variable and the output, but rather, the controller strategy in place. Since the model does not represent the underlying process phenomena, it is immediately invalid once the structure of the sampled process changes (either through physical or operational changes). This problem is similar to the one encountered in identifying non-parametric dynamic models from data generated under feedback control using open-loop methods and assumptions (Box and MacGregor 1974) and was more recently demonstrated by Kresta, MacGregor and Marlin (1991a) for the case of a simulated extraction column.

4.6.7.3 Semi-Confounding of Effects

Colinearity amongst variables works advantageously with PCA and PLS to stabilize the latent variables. However, as with the case of controlled variables, it makes determination of an individual variable's effect on the Y space (based on its regression coefficients) difficult. This is known as semi-confounding of effects. The only conclusion that one can safely draw from the loadings on the variables or biased regression coefficients is that large values indicate the corresponding variable plays an important role (rightly or wrongly) in the model, or is highly correlated with another variable which is important. This issue is discussed in greater detail in section 4.9.

4.6.7.4 Drawing False Causal Relationships

One also has to be careful not to draw direct conclusions as to causality in a model simply because of correlation between certain x variables and the Y space. It is possible that another factor or "underlying" variable which was not collected in the data set may have caused the changes in both the X and Y spaces, leading to the high correlation. Caution must always be applied if one attempts to interpret a variable's loading.

Despite all these drawbacks, a biased regression model developed from historical data can be used for prediction provided that the system modeled does not change physically and continues to be operated in the same fashion as when the data were collected.

4.7 Model Validation and Interpretation Tools

The first two tools used to analyze the results focus on determining the statistical significance of the LVs extracted.

4.7.1 Overall CSV/SD

This value arises from the cross-validation process (as discussed in section 4.5.1) and indicates the statistical significance of each LV. If the value of CSV/SD is less than unity, the latent variable or component has modeled some of the variability in the data set, and has reduced the sum of squared error remaining. Thus, it is considered significant and one proceeds in calculating the next dimension. If, however, the value of CSV/SD is greater than one, the latent variable has modeled more variability than is actually present; it is not significant and the extraction process is stopped. This statistic is used as part of the model dimension cut-off criteria.

Individual CSV/SD values are also calculated for each variable in the data block for PCA and in the Y space for PLS. They indicate which variables are most strongly predicted by the latent variables.

4.7.2 Percentage Variance Explained by Model

For each LV extracted, the percentage variance explained in each block with and without cross-validation is also calculated. Since cross-validation involves actually predicting a portion of the data, it provides a more realistic indication of model fit than the non-cross-validated statistics (i.e., it is conservative and more representative of the prediction power of the LV than the ordinary case).

$$\%DiffSSX = \frac{\%ordinarySSX - \%cross - validatedSSX}{\%cross - validatedSSX} * 100 \tag{4.27}$$

Monitoring the difference between the percentage variance explained using cross-validation and the variance explained for the un-validated procedure draws attention to the percentage of error being predicted. As the difference between these two values becomes greater, the amount of error being modelled is increasing.

4.7.3 Predictive Model Statistics (R^2, Confidence Intervals)

Two statistics were calculated for the predictive models to check for their fit; the correlation coefficient R^2, and 95% confidence limits.

The correlation coefficient value, calculated as

$$R^2 = \frac{\Sigma_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\Sigma_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{4.28}$$

where

$\hat{Y}_i$ = predicted value of Y for sample i

$\overline{Y}$ = mean value for Y

Y_i = observed value of Y for sample i

n = number of samples

measures the proportion of total variation about the mean $\overline{Y}$ explained by the regression model. A high value does not necessarily imply causation but does indicate a good fit. Conversely, a low correlation does not necessarily mean no relationship exists; strong non-linear relationships between Y and X are likely to yield poor R^2 values because of the linear nature of this test.

The confidence intervals for the models were determined from the standard deviations of the fitted prediction errors. Since the expected value of the error mean is zero, the variance of the fitted prediction error can be calculated as follows:

$$s^2 = \frac{\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-1} = \frac{\Sigma_{i=1}^{n}(e)^2}{n-1} \tag{4.29}$$

where

s^2 = variance of the fitted prediction error

e = model error

For 2σ limits, the confidence intervals are calculated simply as

$$C.I. = \overset{+}{-} 2*\sqrt{s^2} \tag{4.30}$$

4.8 Interpreting the Low Dimensional Spaces and SPC Models

Statistical significance, however, does not ensure chemical significance. As the size of a data set is increased (through the addition of samples or measured x or y variables) minute regularities may cause the number of significant latent variables to increase (Wold et al. 1984). Also, as the signal-to-noise ratio decreases, it is hard for PCA and PLS to differentiate between non-linearities in the process and noise. The contribution of later LVs to the model resolution may be so unimportant that leaving them out does not change the results or conclusions drawn from the model. (Wold et al. 1984).

Chemical validation is important to ensure that the model indeed makes sense and to detect possible shortcomings or limitations (Kvalheim and Karstang 1989). PCA and PLS do not assume anything about the information content in the data and can only recognize what is actually in the data. It is quite possible that an analysis would yield no valuable information about the process at all which is merely a reflection of the information content of the data.

An important assumption in these analyses is that the first LVs or dimensions generated (corresponding to the largest eigenvalues of the data set) contain the most useful information relating to the problem and that later LVs describe mostly noise. This must be critically examined in light of the purpose of the analysis.

The latent variables are not always easy to interpret beyond the first or second because each is constrained to be perpendicular to the preceding ones and the process changes and events do not necessarily behave in this manner. However, PCA and PLS may yield good predictive models even though physical interpretations cannot be found for all the LVs.

4.8.1 Inspection of Plots

A key assumption in pattern recognition techniques is that closeness of samples in the pattern space and similarity are related. The distance between samples of a similar class is much smaller when compared to the distance between samples that belong to different classes (Wold et al. 1983). Samples can be visually sorted as belonging to one of the modelled classes (or not belonging to any class). It is also quite possible that one class is easily defined and modeled whereas the other does not appear to be homogeneous with any inherent similarity. Such a class structure is referred to as "asymmetric".

One of the most valuable aspects of the PCA and PLS analyses is the ability to display their results in low dimensional spaces (mostly as two dimensional plots) where a wide range of relationships can be displayed.

Knowledge about what happened during data collection is invaluable for the interpretation of these plots and can be used in two ways. One can list events of interest (such as feed temperature, feed rate and feed quality changes) and locate these on the plots, or one can identify groups of samples at opposing ends of a latent variable's direction and try to determine why they are oppositely related. The sooner the inspection and analysis of the data is performed, the more likely plausible relationships and causes can be identified.

T versus T Plots

Plotting the scores or values of the LVs for each sample can reveal subgroups or classes within the data set and also identify outliers or dynamic periods. These plots indicate what events the latent variables are modeling. Tight clusters suggest major differences between groups are being modeled while loosely packed points suggest smaller continuous variations are being modelled.

P versus P Plots

For PCA data sets with a small number of variables, the P versus P plots reveal the relationship amongst the variables and indicate which ones are most strongly modelled. A variable which dominates a vector may have a large portion of its variation being explained, but if its individual CSV/SD value is near 1.0 (indicating small predictability of this variation) it may indicate modelling of noise.

For data sets with a large number of variables, the P plots become difficult to interpret and other means for analyzing loadings must be used. This issue is addressed under "Normal Plots" below.

W versus W Plots

For PLS analyses, these plots reveal the relationship amongst the X process variables much like the P plots do for PCA. The w vectors, however, are orthogonal, whereas the p vectors are not. If the number of process variables is large, W plots are difficult to interpret and other means of examining the loadings are necessary. See the discussion of normal plots below.

Q versus Q Plots

Q vectors are the loading vectors of the Y space which indicate those Y variables that the LVs are modelling. They help relate changes in the T planes with production changes which are most easily seen in these Q plots.

For data sets with few process variables in the X and Y spaces, the Q plots can be overlaid with W versus W plots to yield one plane where the dominant X process variables and modelled Y variables of each LV can be observed at once.

T versus U Plots

Plotting these two vectors against each other describes the inner relationship being modeled between the X and Y spaces. These plots also show if the relationship is a linear or non-linear one (and therefore how adequately the dimension is able to

describe the relationship) and can also indicate if either the X or Y space is being predominantly modeled. As the slope of the inner relationship line increases, the influence of the X space drops off and the variance in the Y space dominates the model. This suggests that emphasis for interpretation should be placed on the Q plane rather then the T plane. Samples which have a strong influence on the direction of the latent variables are also easily seen in this type of plot.

Normal Plots

Plotting the elements of the vectors for each latent variable on normal paper can provide a visual sorting of the large positive and large negative loadings and thus which variables are dominating the LV model. The shape of the plots can also suggest information about the distribution and relative size of change modelled by the LV. When loadings greater than their mean plot above the normal distribution line and loadings less than the mean plot below the normal distribution line, this suggests that two clusters or distributions exist in the loadings, probably the result of a large change in the process.

Caution must be advised if these plots are used, though, because the plots assume the loading values are independent and identically distributed with fixed mean and variance. For a highly colinear data set, these conditions will not be true. This idea is illustrated in the first PCA analysis in section 5.2.1 for vector p_1 but it not pursued further.

4.8.2 Sum of Squared Prediction Error (SPE) Values

To gauge the modelling error in a monitoring application, the LVs selected for monitoring are used to predict the reference data. The squares of the prediction error for each Y product of a single sample are summed to give the SPE value (sum of squared prediction error).

Since the SPE values should be random in magnitude and independently distributed, samples with abnormally large SPE values warrant further examination.

Only the vectors selected for the monitoring space are used to calculate the SPE values; using more LVs will mean disturbances affecting the unmonitored LVs will not be observable in the SPE plot (Kresta, MacGregor and Marlin 1991a). When test data is introduced to the monitoring space, their SPE values are compared to the 95% confidence level of the reference data to aid in detection of abnormalities.

4.9 Analysis of Loading Vectors and Regression Coefficients

Since the latent variables summarize the variance of the data being studied, it would be beneficial to find out which measured variables play a key role in determining the direction and modelling power of the latent variables. One's first inclination might be to study the loadings or coefficients (in the case of regression models) applied to each variable and assume that large weights indicate important variables and small weights indicate the opposite. There is, however, a problem with this approach.

With PCA and PLS, X variables which are highly correlated amongst themselves and which are all valid predictors of Y will have to "share the weight" of their loadings in the analysis. MacGregor, Marlin, and Kresta (1991b) illustrate this point with the following example:

Given one y variable and five x variables which are perfectly colinear and all equal predictors of y such that the true underlying model is

$$y = 1.0x_i \tag{4.31}$$

where

i=1, 2, 3, 4 or 5,

if all five x variables are used in the PLS analysis, the resultant model will be:

$$y = 0.2x_1 + 0.2x_2 + 0.2x_3 + 0.2x_4 + 0.2x_5 \tag{4.32}$$

Despite the fact that any one x may perfectly predict the y variable, the size of their individual coefficients in the model imply that none of them are extremely important. A key operating variable whose measurement is replicated or highly correlated to other measurements in the process data set (e.g., a regenerator temperature which is measured at different locations in a bed) may not yield a significantly large weight or coefficient despite the well known theoretical and practical importance of such a variable.

Although process variables will not be perfectly colinear, they often contain a high degree of correlation and, as the example shows, attempting to draw conclusions about the importance of variables simply from their loadings or coefficients can be very dubious.

The only conclusion that can be drawn is that variables with large loadings or coefficients play an important role in the model or are correlated to another variable (or variables) which are important. The only way to draw relationships between cause and effect in regression equations is to use designed experiments for data collection or to use mechanistic models (MacGregor, Marlin, and Kresta 1991b).

The issues discussed in this chapter are addressed specifically with respect to the industrial data in chapter 5.

CHAPTER 5: DATA PRETREATMENT AND PRELIMINARY PCA

This chapter outlines the data collection, variable selection, scaling and pretreatment performed on the process data before any analyses were conducted. It also contains the results of the preliminary analysis using PCA on the X and Y spaces (separately) of the full data set (1170 samples), followed by a second PC analysis using a subset of data (approximately eleven days' worth of hourly averages).

5.1 Data Collection and Pretreatment

5.1.1 Collection of Process Data

The data used in the following analyses were downloaded directly from the FCCU process computerized database to LOTUS 1-2-3 spreadsheets. The aim was to collect informative yet steady-state data; since the residence time of the products in the unit is one hour or less, the process engineers felt that hourly data would best represent steady-state operations. Data of a shorter time interval (a few minutes) was also available but this was expected to be highly dynamic, requiring many latent variables for proper modelling and making it difficult to interpret the model. Daily averages were also available, but these values were expected to show little of the type of behaviour that could potentially be unveiled (feed quality changes, mechanical problems, short upsets, etc.). The data collection took place over a three and a half month period during the

autumn of 1990.

Additional off-line information available included; product analyses done at the unit site twice daily, lab analyses on feed and product streams daily, and catalyst analyses done several times a week. Due to the large amounts of missing data in these sets (compared to the hourly readings available for the process conditions) they were not used directly in the PCA or PLS analyses. They were, however, retained for their usefulness in interpreting the types of events taking place within sample clusters (unit operational moves, feed quality changes, disturbances, etc.)

5.1.2 Variable Selection: What is X and What is Y?

The most important consideration for selection of process variables for the X and Y spaces was the purpose of the model. Eleven volume yields and selectivity values were selected to form the Y space and are listed in table 5.1.

The percent volume yields are based on the unit feed rate; products 9, 10 and 11 are functions of some of the other eight, generating a somewhat correlated Y space.

Using all available x variables (over 300) for the analyses posed a substantial computational problem, requiring far too much time for model building or prediction. Although using only a subset prejudges a variable's importance, using fewer than 300 x variables reduces the noise in the data set and allows all 1400 plus samples to be run at once for the initial look at the plant data. The set was thus pared down to 136 X variables by having the process engineers select key variables and by including any extra manipulated or controlled variables in the set. The product flow rates (upon which the volume yields are based) were not included in the X space because they would have correlated very strongly with the percent volume yields and, in so doing, might have masked subtler effects of other x variables.

Table 5.1--Y Space Product Variables

Number	Product	Units
1	Fuel (dry) gas yield	% volume gas/volume feed
2	Propane yield	% volume
3	Butane yield	% volume
4	Gasoline yield	% volume
5	Light Cycle Oil yield (LGO)	% volume
6	Intermediate Cycle Oil yield (IGO)	% volume
7	Heavy Cycle Oil yield (C.O.)	% volume
8	Coke yield	weight of coke per volume of feed
9	Liquid products yield (a function of 2,3,4,5,6 and 7).	% volume
10	Gasoline selectivity (a function of 4)	volume/volume
11	Coke selectivity (a function of 8)	volume/volume

Selection of the sample space was dependent upon the purpose of the analysis (as shown in table 1.1 in Chapter 1) and is discussed in the introduction of each analysis.

5.1.3 Assumptions about the Data

Several assumptions were made concerning the data collected. Since PCA and PLS are modelling the correlational structure of the data, a key assumption was that this structure does not change over the time history being studied and will be the same in any future data used with the model. If the plant was operated in different fashions at different

times then this will not be the case and the percentage of cross-validated sum of squares explained by many latent variables will be much lower than their percentage of fitted sum of squares.

Another assumption was that changes in the process variables caused by disturbances or operational moves were much larger than instrument error (or noise). This was not tested directly, however, PCA and PLS should sort these types of variables out. If the noise of a variable is much larger than its effect on a process output, that process variable should not contribute heavily to the latent variables. The noise, however, will contribute to the overall variability in the data matrix (or matrices), thus slow down the explanation of matrix variance. Scaling plays an important role with respect to this issue.

It was also assumed that the hourly averages represent steady-state. Auto- and cross-correlations were performed on all the variables to test how close the data fit this assumption and whether or not data shifting would be necessary. Details of this are contained in Appendix A. The light gas products did not show any significant lags while the heaviest liquid products appeared to be correlated to samples two hours earlier. Intermediate products showed a scattering of correlations lying between these two ranges.

This suggested that time shifting of the data might be necessary, so to test this, samples from the feed, reactor, regenerators, air blower and desuperheater were replicated and shifted back in time by one hour. This matrix was then added onto the original X space of 136 variables to form an augmented X block of 221 variables, as shown in figure 5.1. Two PLS runs were then conducted; one using the original X (1469 samples by 136 variables) and Y spaces (1469 samples by 11 variables), the second using the augmented X space (1450 x 221) and original Y space (1450 x 11).

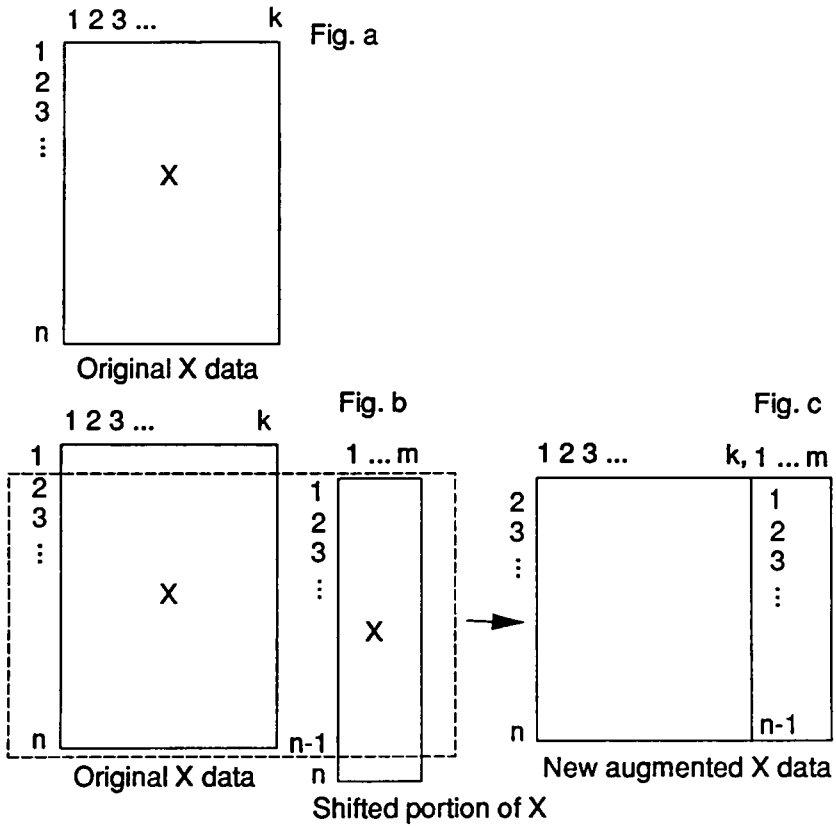


Figure 5.1: Building a time-shifted X data matrix: a) original data block, b) a portion of the X data is replicated and shifted back one time unit to represent the previous time period, c) new augmented X data block.

It was expected that, if time shifting of the data was beneficial, the second PLS run should describe more of the Y space within the first few latent variables than the first run. In fact, both runs showed similar descriptive power in the first ten latent variables. This suggested that the time shifted (n-1) data points were highly correlated to the

present n samples and that few radical dynamic events were occurring (where the predictive power of the n-1 samples should be evident in boosting the percent variance explained in the Y space).

Since no real improvement in the modelling of Y was found with the augmented X matrix, it was felt that time shifting of the data sets would not be necessary.

A high degree of confounding was also expected within the data set amongst key operating variables due to the manner in which the FCCU was run. For instance, asphalt production mode (referred to as a "boomer" run) is characterized by cooler , higher quality FCCU feeds. Feed and catalyst qualities, which are not measured directly, are known to change substantially throughout the data set, and there was a known drift in the operations which was confounded with time.

5.1.4 Scaling

As pointed out in Chapter 4, scaling plays a key role in the PCA and PLS analyses. Auto-scaling was used for all the analyses so as not to impose a pre-conceived structure on the data. Section 6.5 considers how the results of the analyses could be examined for effects of scaling.

5.1.5 Extreme Outliers

In the first run through, 1469 samples were used. Examination of both the PCA results on the X space and the PLS results (namely the T versus T plots and U versus T plots) quickly revealed the presence of transient data (i.e., samples which were part of transitions between different operating modes and hourly averages leading into or out of unit shutdowns). These were dominating the later latent variables (from the seventh

onwards) and might have been masking smaller changes in the data. Transient data was verified from operating records then removed from the data set, leaving 1170 sample periods, for the subsequent analyses.

5.2 Preliminary Analysis Using PCA

Since the purpose of the preliminary analyses was to inspect the data set, all 1170 hourly average readings on 136 x variables and 11 y variables were used. Table 5.2 lists the process events known to have occurred during the time period presented by this data and table 5.3 notes when breaks in the time history occurred. This information was used later to help interpret sample clusters, abrupt changes and the meaning of the latent variables.

Note that the terms latent variable and principal component can both be (and are) used for PCA.

Table 5.2–Process Events in PCA and PLS in X and Y Spaces (X=1170 x 136) (Y=1170 x11)

Event	Sample Number	Event	Sample Number
Low Ni and V	1-7, 389-482, 767-801	Low catalyst MAT	...460-642... ...696-777...
High Ni and V	556-578...	High unit feed rate	1-388 389-762 767-936
Moderate Ni and High V	824-857	Medium unit feed rate	937-1111
High Ni and Moderate V	1112-1170	Low unit feed rate	1112-1170
Boomer (asphalt-producing mode)	1-7 389-427 767-823	High feed temperature	179-184 430-510
Non-boomer (normal refinery feed)	8-388 428-766 824-1111 1112-1170	Low feed temperature	1-7 398-427 767-823 860-981 1033-1066 1071-1170
High catalyst MAT	....66 ...670-695		

Legend:
- ...# = the period starts leading up to this sample number
- #... = the period dies out near this sample number
- Ni = nickel
- V = vanadium
- MAT = (catalyst) micro-activity test

Table 5.3--Time Breaks in the PCA and PLS Data Set

Sample Sets	Dates	Break Between Sets (Hours)
1-7	Oct 1 8:00 - Oct 1 15:00	7
8-160	Oct 1 22:00 - Oct 10 9:00	3
161-277	Oct 10 12:00 - Oct 15 8:00	5
278-322	Oct 15 13:00 - Oct 17 9:00	48
323-388	Oct 19 9:00 - Oct 22 3:00	41
389-427	Oct 24 20:00 - Oct 26 10:00	12
428-556	Oct 26 22:00 - Nov 1 8:00	25
557-642	Nov 2 9:00 - Nov 5 22:00	3
643-695	Nov 6 1:00 - Nov 8 8:00	26
696-766	Nov 9 10:00 - Nov 12 8:00	60
767-803	Nov 14 20:00 - Nov 16 8:00	8
804-823	Nov 16 16:00 - Nov 17 11:00	35
824-857	Nov 18 22:00 - Nov 20 8:00	241
858-1025	Nov 30 9:00 - Dec 7 8:00	32
1026-1066	Dec 8 16:00 - Dec 10 8:00	124
1067-1111	Dec 15 12:00 - Dec 17 8:00	657
1112-1170	Jan 14 17:00 - Jan 17 3:00	

5.2.1 PCA Analysis of X

The goals of the X PCA analysis were to identify: interesting periods in the data such as specific clusters or subgroups, abrupt changes in the "location" of the operations, and subtle changes in the data (expected to be revealed by smaller or later latent variables). The analysis would also give an estimate of the number of dimensions needed to describe the X space (for PLS work). Table 5.4 contains the statistical results of the analysis.

Table 5.4--Statistical Results from PCA of X (X=1170 x 136)

LV	Ordinary % SS X	Cumul. % SS X	Cross-validated % SS X	Cumul. % SS X	% Diff SS X*	Overall CSV/SD
1	27.6	27.6	20.0	20.0	38.0	0.894
2	13.8	41.4	9.7	29.7	42.3	.930
3	12.3	53.7	8.9	38.6	38.2	.920
4	5.9	59.6	3.7	42.3	59.5	.959
5	5.0	64.6	3.4	45.7	47.1	.956
6	4.1	68.7	2.6	48.3	57.7	.963
7	3.0	71.7	1.7	50.0	76.5	.972
8	2.6	74.3	1.6	51.6	62.5	.971
9	2.2	76.5	1.2	52.8	83.3	.977
10	2.0	78.5	1.0	53.8	100.0	.977
11	1.8	80.3	0.7	54.5	157.1	.983
12	1.4	81.7	0.3	54.8	366.7	.992
13	1.4	83.1	0.7	55.5	100.0	.980
14	1.3	84.4	0.6	56.1	116.7	.981
15	1.2	85.6	0.6	56.7	100.0	0.980

$$*\%DiffSSX = \frac{(\%ordinarySSXexplained - \%cross - validatedSSXexplained)}{(\%cross - validatedSSXexplained)} * 100$$

Table 5.4 shows that the difference between cross-validated and ordinary sum of squares (SS) explained in X was quite large for most vectors and indicated that the predictive model fit was not tight. This may be due to the presence of substantial non-linearities amongst and within sample clusters and the large number of clusters (or sub-groups) present which make fitting to a linear model difficult. Confounding amongst the samples (i.e., having a sample belong to more than one group) may have also contributed to this.

Although all the CSV/SD values for the first fifteen latent variables were significant, their modelling power (exemplified by the cross-validated percentage variance explained) was quite small by the eleventh LV. The first ten LVs accounted for 53.8% of the cross-validated SS in X while LVs eleven to fifteen modelled only an additional 2.9%.

The gap between the ordinary and cross-validated results also widened for these later LVs. Thus, the percentage difference SS in X value can aid in making a subjective decision on what LVs to leave out of the model.

Next, the T plots of figure 5.2a-g were examined for trends and clusters in the samples. Note that in these figures, the T vectors were plotted against T1, the first and most explanatory vector, for the reason of simplicity. In many cases, however, it may be more meaningful to use different combinations of the vectors to provide more interpretable planes. This will depend upon the purpose of the analysis and is considered in more detail in the monitoring space section 7.1.3. Table 5.5 provides details on the sample clusters which were revealed by the PCA analysis.

Due to the large number of process variables present in the X block, the P planes pose a more challenging interpretation problem. Figure 5.3, where P1 and P2 are plotted against each other, illustrates this. The variables' loadings fill a single region on the plane and cannot be easily sorted visually to identify the most influential for the individual vectors. The plane does, however, nicely display to what extent a variable contributes to the explanation of both directions at the same time.

An easier way to examine the loadings would be to display them as a bar graph or use a normal plot. Figure 5.4 shows the normal plot for vector P1. The shape of the points on the plot suggest a non-normal (i.e., a multimodal) distribution for the loadings. This could be interpreted as meaning that the change in operations being described by the first LV is so great, there really exist several different modes of operation.

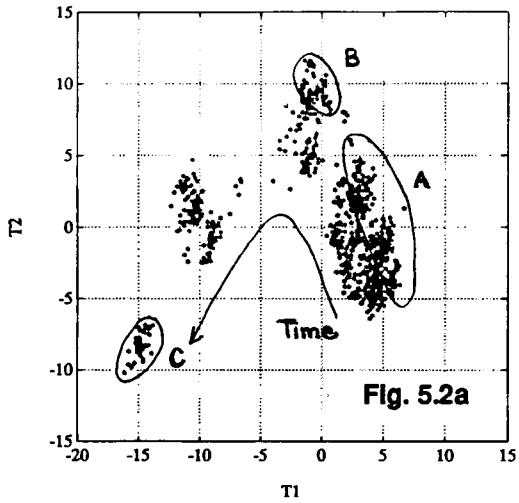


Fig. 5.2a

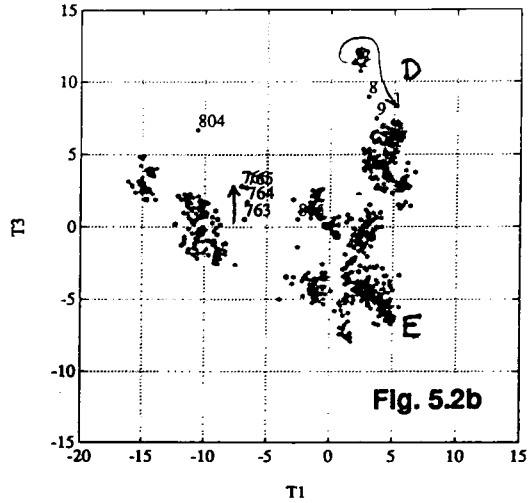


Fig. 5.2b

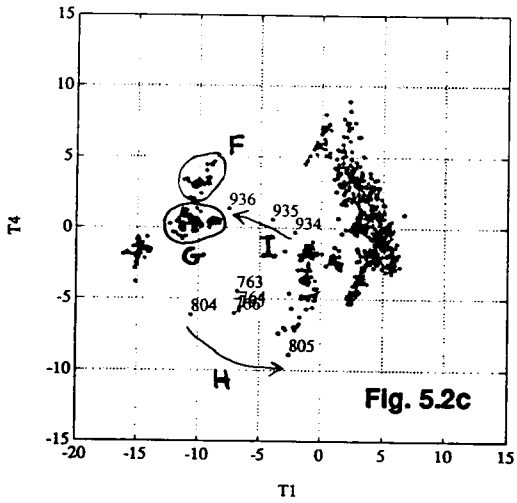


Fig. 5.2c

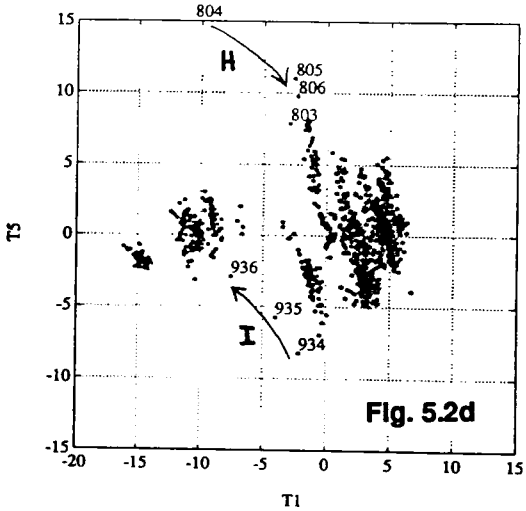


Fig. 5.2d

Figures 5.2a-d: T (sample) planes from PCA of X.

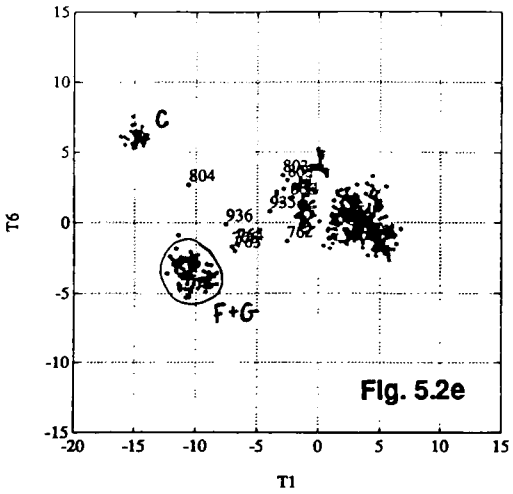


Fig. 5.2e

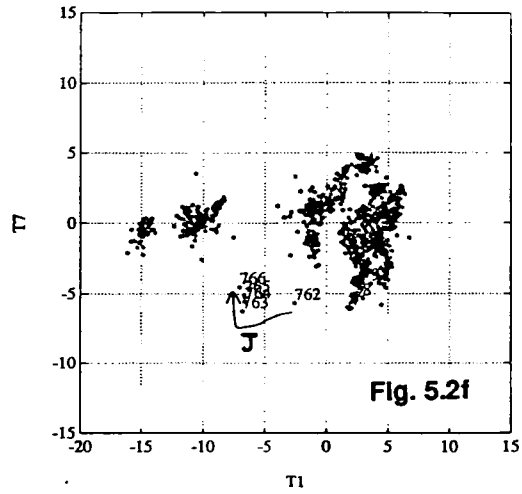


Fig. 5.2f

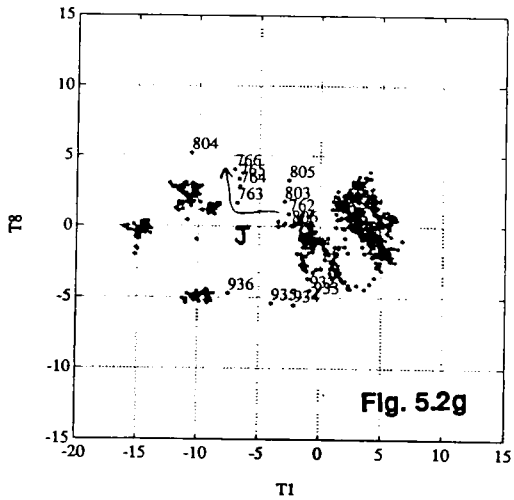


Fig. 5.2g

Figures 5.2e-g: T (sample) planes from PCA of X.

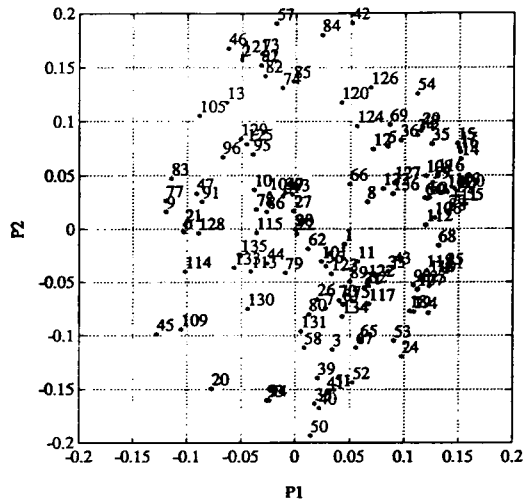


Figure 5.3: P1-P2 plane from PCA of X.

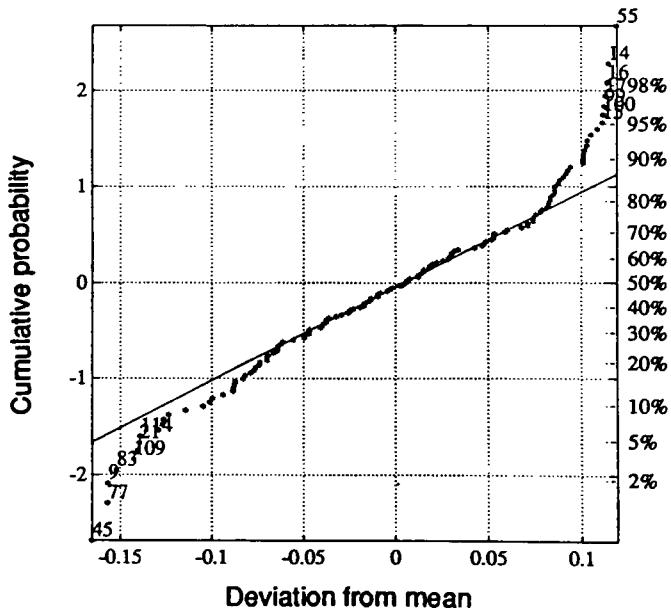


Figure 5.4: Normal plot of P1 loadings from PCA of X.

Due to the underlying assumptions involved in using the normal plot, caution is advised when applying this technique to the loadings. A bar graph of the loadings would provide much the same information about which variables have very large absolute loading values without assuming an underlying distribution. This latter approach was used to pinpoint some of the variables which contribute heavily to each LV in the models and which are discussed in table 5.5.

Table 5.5--Analysis of T and P Plots from PCA of X

Sample Space	Variable Space
T1: A = 1-388: Non-boomer, fairly typical operations, high feed rate. B = 398-427,767-823: Boomer, high feed rate, and low metals. C = 1112-1170: Non-boomer, low feed rate, and high nickel content. There is a distinct time trend which is confounded with a known drift in operations plus a production throughput decrease. The four distinct groups with no real outliers suggest that the vector is describing major changes within the data set.	**P1:** Several controlled variables (CVs) have large loadings, suggesting that T1 is describing some operational moves. These are: - feed flow to the nozzles, - excess oxygen in the flue gas, - second regenerator pressure, - duty and bottom temperature of the desuperheater, - total duty of LGO, and - debutanizer feed flow. Manipulated variables (MVs) with large loadings may simply reflect moves made to maintain CV set points but could also indicate that disturbances in the system are being modeled. - flow of stripping steam, - pressure difference across lift air valve - liquid flow to the rectified absorber - gasoline flow to the rectified absorber
T2: This vector differentiates group B from C. B represents low gravity, low metal feed quality at a high throughput rate while C represents high gravity, high nickel content feed at a low throughput.	**P2:** The largest loadings are for the first and second regenerator temperatures and the C/O ratio based on the second regenerator flue gas composition.

Table 5.5-- Continued

Sample Space	Variable Space
T3: D = 1-9: Low vanadium metal feed, tail end of a boomer mode. E = 556-570, 800-823, 824-858: Moderate to high vanadium content in feed. Samples 10-858 were originally in one group (see T1 versus T2 plot) but here are split up into about six subgroups. Latent variable appears to distinguish between low and high vanadium feed quality in samples (D versus E respectively).	**P3:** Large loadings are found for: - flow of stripping steam, - first regenerator dense and dilute bed temperatures, - flow of the main fractionator top reflux, and - temperature of stream exiting the debutanizer reboiler
T4: F = 936-1066 G = 1067-1111 Samples 1066 and 1067 are five days apart; groups F and G represent the change in operating point over this time, which may be due to a decrease in total feed flow. Samples 1-858 (A,B,D and E) are less distinct in their groupings.	**P4:** Large loadings appear for mostly MVs based at the desuperheater and fractionator.
T5: H = 804-806: Transition out of a unit shutdown. I = 934-936: Dec 3rd 13-15:00 Feed rate dropped by 8%. The latent vector is modeling smaller but significant events, a transition from a shutdown (H) back to normal operations and a drop in feed rate (I).	**P5:** Variables with large loadings are a mixture of MVs and CVs at the desuperheater and MVs at the debut and depropanizer.
T6: Latent variable very strongly separates the combined group of F+G from C (the most recent data). There is a one month lag between these two groups. Interpretation is questionable but suspect it to be operational changes.	**P6:** Key variables are a mix of minor flows at the feed manifold, and indicators of feed quality, plus a handful from desuperheater to the rectified absorber.

Table 5.5-- Continued

Sample Space	Variable Space
T7: J = 762-766: Preparation for boomer. Major dynamics (moving towards and out of a boomer operating mode) are being modeled; most points cluster around the zero axis (i.e., within a score of +-6).	**P7:** -fractionation variables
T8: More transient moves are being modeled while remaining data is centered at zero point on axis.	**P8:** - feed temperature, - riser outlet temperature, - temperature of air to both regenerators - lower and intermediate section temperatures and flows at main fractionator - rectified absorber sponge oil circuit
T9: (not shown) The transition out of a unit shutdown is distinguished from the rest of the data samples.	**P9:** (not shown) - runback and HFD from crude unit - air temperatures exiting blower and to both regenerators - second regenerator main air flow and catalyst level
T10: (not shown) K = 672-676: Unknown transient. Modeling unknown variation.	**P10:** (not shown) - runback flows, - dispersion steam total and to each nozzle (which are a function of feed quality and thus highly correlated to runback flow) - catalyst levels in both regenerators - fractionator IGO and upper circulating refluxes

Latent variables eleven to fifteen appeared to model only a handful of radical points and thus were not shown.

5.2.1.1 Summary of PCA of X

In general, PCA illustrated that the FCCU during the time period studied had many operating "windows" or many different areas in the T planes where the process operated. The process was not operated stably at a fixed point, but was moving continually. In a quality control situation, this would not occur.

Latent variable decomposition did not provide components that related directly or clearly to key variables or known phenomena. Such a result cannot be expected, however, unless the data collection has been designed to yield an orthogonal separation of phenomena. By using information about what occurred with the process during data collection, some rough interpretations were made.

The difference between the variance explained by cross-validation and that by ordinary means suggested that only the first ten dimensions modeled significant process variability in the X space. These LVs accounted for 53.8% of the cross-validated sum of squares (SS) in X. Nearly half of the process variance was not accounted for in the model.

The first few latent variables described major steady-state locations or operating windows of the FCCU, while later vectors accounted for dynamic changes. The first and second latent variables separated the samples into four distinct groups. The first latent variable appeared to be describing throughput effect, however, this was also confounded with time and a known drift in the operations.

Latent variables two and three appeared to be related to feed quality (specifically metals content) which has a poisoning effect on catalyst and again, there was confounding of feed quality with catalyst quality (although the time constant of one can be quite different from the other). Without having designed the data collection though, it was difficult to say with certainty that these were the only phenomena that

these two LVs described.

The roles of latent variables four and six could not be determined, although distinct group separations were evident in the T plots. The fifth vector appeared to model two distinct transients in the data. More dynamics, transients and other abnormal operations (e.g., start-up) were modeled by the later dimensions seven to ten; this was quite evident from the T plane plots.

The remaining LVs eleven to fifteen had poor model prediction statistics and their T plane plots suggested that they were describing noise. Together, they only explained 2.9% of the sum of squares X so in all likelihood, they could be dropped from the model without loss of predictive power.

One point discovered after running the PLS case was that in some cases, sub-group separations were a result of changes in the Y space but without this extra information, making this interpretation strictly from the PCA results was next to impossible. This point is discussed in detail under the PLS results analysis section 6.4.

PCA was able to pull the data apart into clear sub-groups and highlight major transients, but due to the great amount of variance still to be explained after ten dimensions, it appeared that the data set may have contained too many deliberate process changes to allow for adequate modelling in a small number of latent variables. Using a sub-set of the data may be more fruitful in this respect.

From the above analysis, one would expect that a PLS analysis performed on the same X space (with an accompanying Y space) would require at least ten vectors to explain major components of the X space.

5.2.2 PCA of Y

The purpose of performing PCA on the Y space alone was to examine the structure within this space for information which would be useful in interpreting the PLS case later. Tables 5.6 and 5.7 list the statistical results of this analysis.

Table 5.6--Statistical Results from PCA of Y (Y=1170 x 11)

LV	Ordinary % SS Y	Cumul. % SS Y	Cross-validated % SS Y	Cumul. % SS Y	% Diff SS Y*	Overall CSV/SD
1	40.5	40.5	26.7	26.7	51.7	0.856
2	20.1	60.6	3.4	30.1	491.2	.971
3	14.0	74.6	7.7	37.8	81.8	.896
4	8.6	83.2	0.8	38.6	975.0	.983
5	7.8	91.0	3.1	41.7	151.6	.904
6	4.7	95.7	0.7	42.4	571.4	.959
7	2.8	98.5	0.8	43.2	250.0	.896
8	1.5	100.0	1.0	44.2	50.0	.559
9	0.0	100.0	0.0	44.2	100.0	.817
10	0.0	100.0	0.0	44.2	-41.2	.505
11	0.0	100.0	0.0	44.2	n/a	0.516

$$*\%DiffSSY = \frac{(\%ordinarySSYexplained - \%cross - validatedSSYexplained)}{(\%cross - validatedSSYexplained)} * 100$$

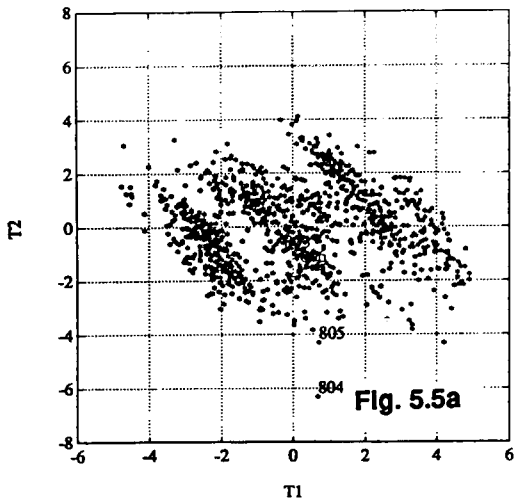
The first eight LVs described 44.2% of the cross-validated SS in Y and although later LVs had significant CSV/SD values, the amount of cross-validated variance they explained was nil. The large discrepancy between ordinary and cross-validated results suggests a high degree of inconsistency in the data.

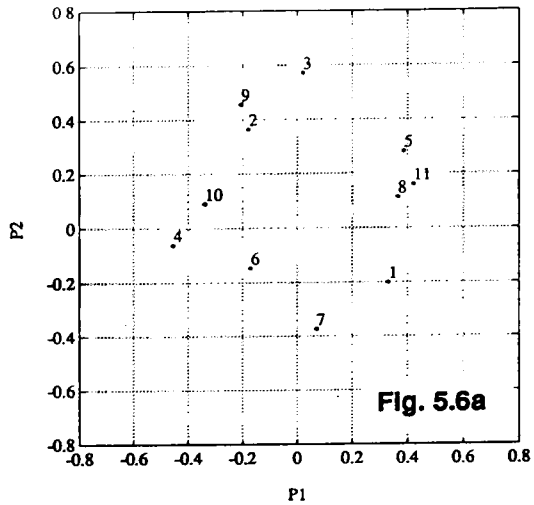
Table 5.7–Y Product CSV/SD Values from PCA of Y

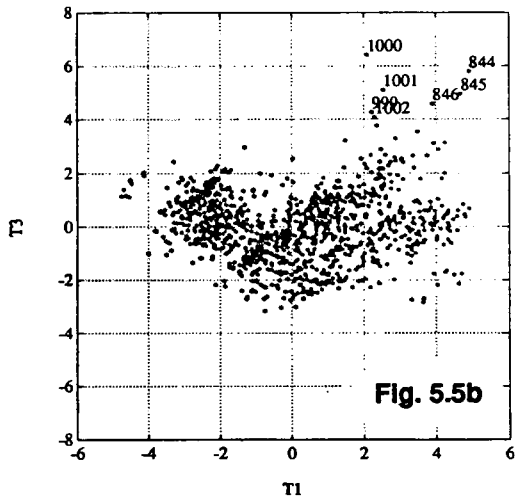
LV	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11
1	0.84	0.95	1.00	0.60	0.75	0.96	1.00	0.76	0.90	0.87	0.69
2	1.02	0.94	0.95	0.92	0.97	0.99	0.99	0.99	0.93	0.99	0.98
3	0.89	0.98	0.88	0.95	0.98	1.00	0.88	0.98	0.78	0.56	0.90
4	1.00	1.00	1.00	0.99	0.98	0.96	1.00	1.00	0.98	1.00	0.79
5	0.95	0.88	0.98	0.94	0.70	1.00	1.00	0.89	0.41	0.94	0.72
6	0.96	1.00	1.00	0.94	1.00	0.86	0.95	1.00	0.99	0.62	0.99
7	1.02	0.75	0.87	0.82	0.56	1.08	0.98	0.97	0.90	1.00	0.94
8	0.61	0.43	0.61	0.54	0.95	0.42	0.61	0.42	0.78	0.58	0.80
9	0.74	1.00	0.98	0.60	0.83	0.44	0.84	0.38	0.98	0.94	0.79
10	0.54	0.47	0.54	0.47	0.54	0.47	0.54	0.47	0.54	0.47	0.54
11	0.57	0.44	0.57	0.44	0.57	0.44	0.57	0.44	0.57	0.44	0.57

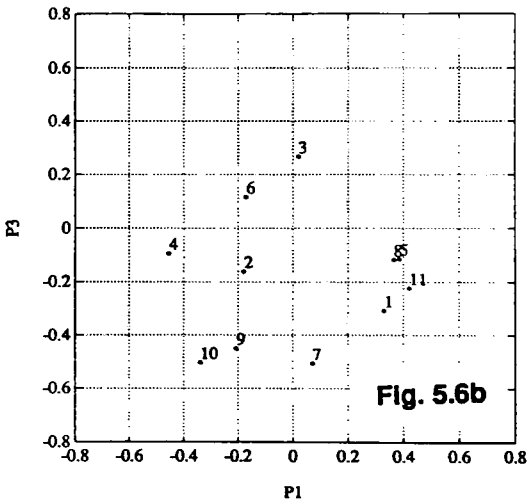
This is quite possible for the Y space studied here since several of the yields and selectivity values are functions of other y variables. Liquid product yield 9, for instance, could conceivably have a consistent value even if the volumetric yields 2,3,4,5,6 and 7 (of which it is a function) changed. In this way, inconsistencies in the relationship amongst the y variables are introduced and make modelling of the data space difficult.

Figures 5.5a-g and 5.6a-g show the T and P planes, respectively, generated from the analysis. Table 5.8 summarizes the major trends observed per component.



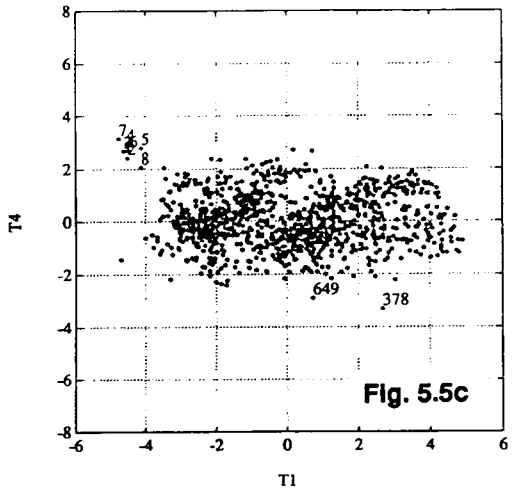


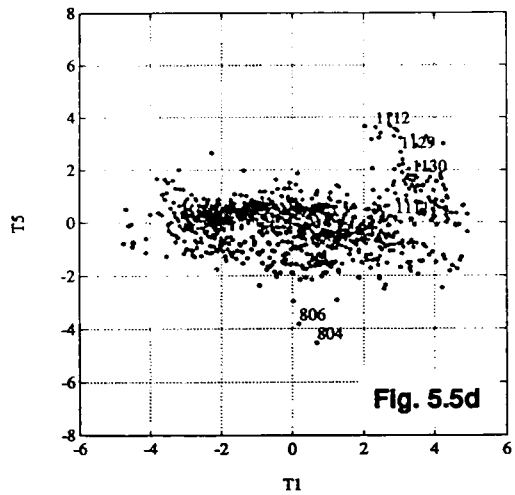




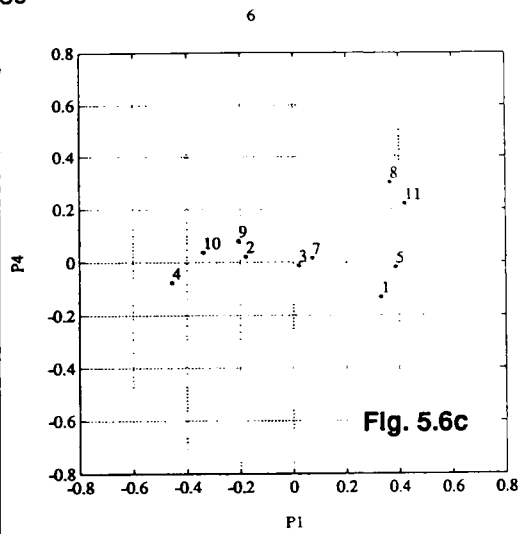
Figures 5.5a-b: T (sample) planes from PCA of Y.

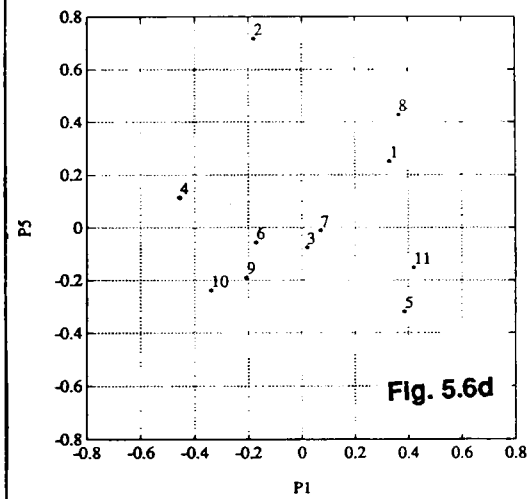
Figures 5.6a-b: P (process variable) planes from PCA of Y.



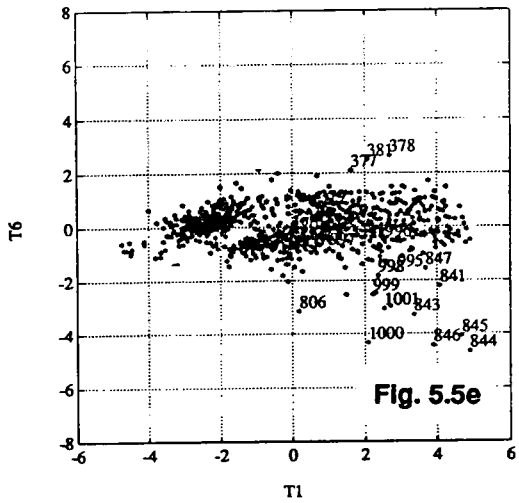


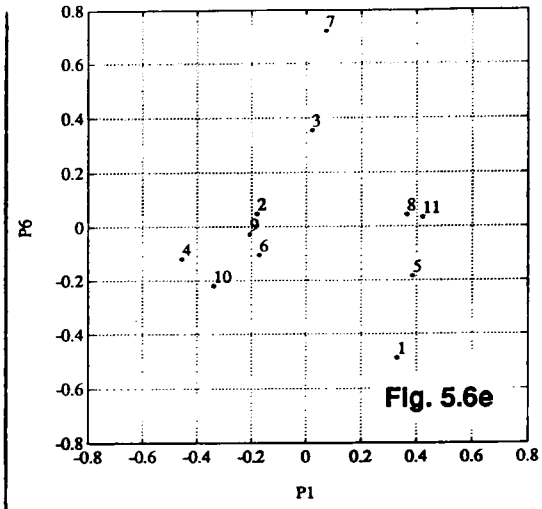
Figures 5.5c-d: T (sample) planes from PCA of Y.

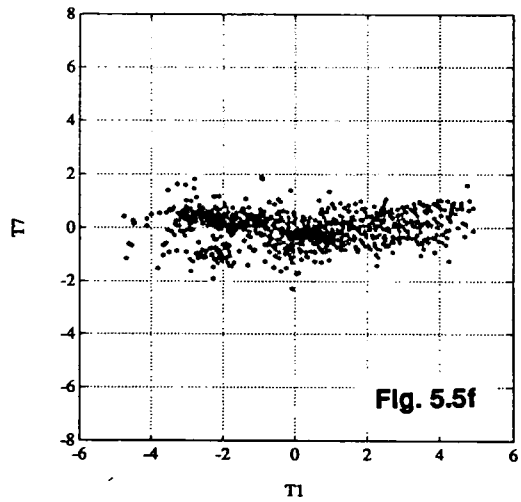


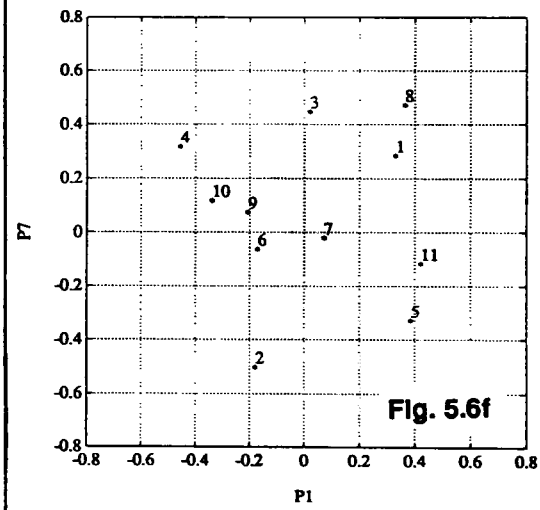


Figures 5.6c-d: P (process variable) planes from PCA of Y.



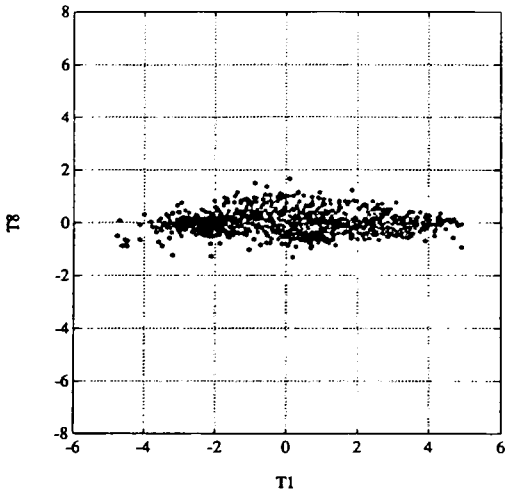




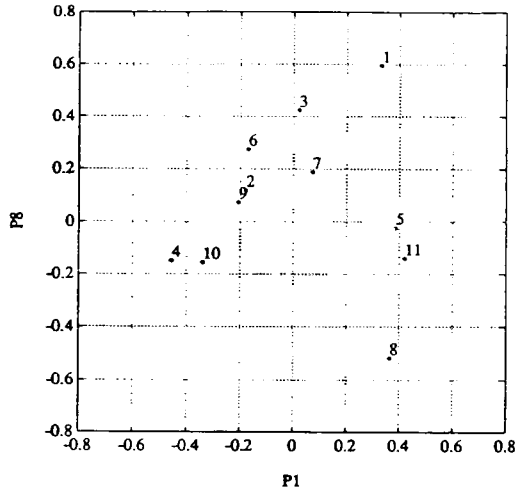


Figures 5.5e-f: T (sample) planes from PCA of Y.

Figures 5.6e-f: P (process variable) planes from PCA of Y.



Figures 5.5g: T (sample) plane from PCA of Y.



Figures 5.6g: P (process variable) plane from PCA of Y.

Table 5.8--Analysis of T and P Plane Plots from PCA of Y

T Plane	P Plane
T1, T2: Points in the T1-T2 plane appear to form three elliptical regions in succession lying parallel to the T1-T2 diagonal, but they are so close together that, overall, the points form a singular cloud; this suggests that the vectors are describing variation common to all samples.	**P1, P2:** The P1 vector models gasoline (4) versus LGO (5) and coke (8 and 11). This is substantiated by the individual CSV/SD values in table 5.7. P2 appears to model undesirables (7) versus desirables (2,3, and 9) although all CSV/SD values for this component are high indicating that modelling power is low and that the component probably represents noise.

Table 5.8--Continued

T Plane	P Plane
T3: This dimension has a few outlying points affecting its direction: 844-846: Cause unknown 849-857: Might be due to air blower problems 995-1001: May represent adjustments to the fractionator top temperature, rectified absorber (RA) temperature, and lean oil rate to the RA.	**P3:** The P1-P3 space shows no striking pattern, yet the CSV/SD values suggest that P3 is modeling liquid, gasoline and butane yields (9,10 and 3) and to a lesser extent fuel gas (1).
T4: This vector pulls start-up points 1 to 9 out slightly from the rest of the data.	**P4:** Interestingly, the P4 direction appears to be heavily influenced by IGO yield (6). However, the PC's overall modelling power is low. The lowest CSV/SD value amongst the Y products (and thus the largest % explained for a Y) is for coke selectivity (11). A smaller component of variance is explained for 11, but it is highly predictable whereas more variation in 6 is being removed although the component may not be as highly predictable.
T5: Separates samples 1113 to 1128 (which coincide with the start up of a feed cooler). It also highlights the start up points 804 to 806.	**P5:** Here, the CSV/SD values and the plot coincide; P5 models LGO, liquid, and propane yields (5,9 and 2) and coke selectivity (11).
T6: This LV is influenced by the very dynamic samples 841 to 847 and 995 to 1001.	**P6:** Here again, the CSV/SD values suggest LGO and gasoline (6 and 10) are being modeled while the plot separates C.O. (7) and fuel gas (1) from the rest.
T7 and onwards: These plots show no distinguishing features in the sample space.	**P7 and onwards:** P7 appears to be modeling the relationship between the light and medium weight products (2 and 3 versus 4 and 5). Remaining dimensions don't appear to describe much structure in the Y space.

5.2.2.1 Summary of PCA of Y

The first component focused on differentiating between the gasoline yield and selectivity (4 and 10) and the undesirable products of dry gas and coke (1,8 and 11); it also focused on modelling the relationship between gasoline yield (4) and LGO yield (5). The second PC appeared to fit the difference between the light and heavy products. The third PC modelled butane yield and gasoline selectivity (which decreased as the butane yield increased). Principal component four appeared to model noise and the IGO yield appeared to contribute significantly to this. Principal component five modeled propane at one end and light cycle oil, coke selectivity and liquid yield at the other; the modelling power for the remaining components was questionable.

The T plane plots showed that despite whatever was happening with the actual process unit, for the most part, the product yields lay within a single cluster in the latent variable space. This contrasted sharply with what was seen in the X space PC analysis where several operating points or regions could be clearly defined. Carry over of these operating point changes from region to region in the X space was not seen to cause several different Y output regions (except in the cases of start-up and major moves).

Alone, the Y analysis was not that informative. The eight basic products required around six to eight dimensions to explain their variance, although leaving a substantial amount of the Y space unexplained. However, the analysis did reveal that the plant was producing its products within a consistent and single output window.

5.2.3 PCA of a Subset of the FCCU Data

The purpose of this analysis was to see if by examining a shorter period of time , one could isolate events which occurred in the process and thereby learn about special causes.

The portion of data selected for further PC analysis was the sample set 428-695 in which significant catalyst poisoning (and corrective action to counter it) took place. In the original PCA work, these samples sat in a region bounded by 0 and +5.0 on the T1 axis and -5 to +5 on the T2 axis; no clear separations amongst these samples were obvious.

Both the X and Y spaces were analyzed using PCA. Known changes in key process variables are listed in tables 5.9 and 5.10. The statistical results are presented in tables 5.11 and 5.12 and the T and P planes for each case are described below.

Table 5.9--Key Process Changes In Subset of FCCU Data

Event	Sample No.
Low Ni and V	1 - 55
High Ni and V	129 - 151
Low MAT	33 - 213
High MAT	243 - 268
High Feed Temperature	1 - 80
Low Feed Temperature	130 onwards

Table 5.10--Crude Diet for Subset of FCCU Data

Start of Diet (Sample No.)	Light (%vol)	Heavy* (%vol)	Other (%vol)
1	96.0	2.0	2.0
51	41.1	57.6	1.3
85	97.9	0.2	1.9
107	75.2	23.8	1.0
136	81.0	18.0	1.0
199	82.7	17.3	0.0

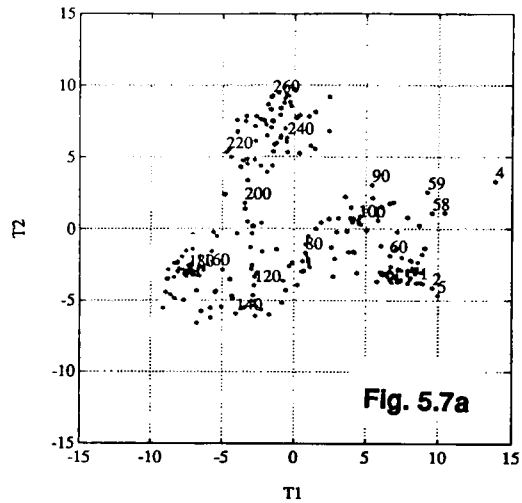
* High metals content typically accompanies the heavier feed stocks.

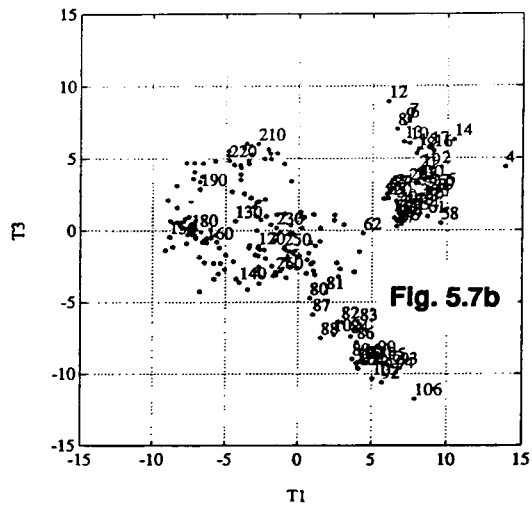
5.2.3.1 PCA of the X Subset

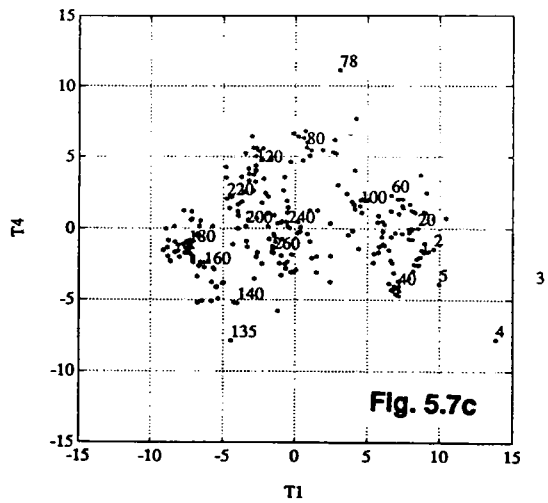
Table 5.11 contains the statistical results of the PC analysis on the short X data set. Although the first twelve principal components had acceptable CSV/SD values, the first four vectors described 53.3% of the cross-validated sum of squares (SS) in the reduced data set and showed the most distinct groupings. Later PCs described decreasing amounts of the SS which focused on distinguishing small handfuls of points from the rest of the samples. Many of the shifts in the samples coincided with changes in the content of the refinery crude unit diet.

Table 5.11–Statistical Results from PCA of X Subset (X=268 x 136)

LV	Ordinary %SS X	Ordinary Cumul. %SS X	Cross-validated %SS X	Cross-validated Cumul. %SS X	% Diff SS X	Overall CSV/SD
1	24.6	24.6	22.6	22.6	8.8	0.879
2	14.9	39.5	13.6	36.2	9.6	0.905
3	12.1	51.6	11.0	47.2	10.0	0.904
4	7.0	58.6	6.1	53.3	14.7	0.934
5	4.7	63.3	3.9	57.2	13.1	0.952
6	3.7	67.0	2.7	59.9	37.0	0.962
7	3.1	70.1	2.3	62.2	34.8	0.964
8	2.8	72.9	1.9	64.1	47.4	0.967
9	2.5	75.4	1.7	65.8	47.0	0.968
10	2.1	77.5	1.3	67.1	61.5	0.972
11	1.7	79.2	0.3	67.4	466.7	0.992
12	1.5	80.7	0.3	67.7	400.0	0.993







Figures 5.7a-c: T (sample) planes from PCA of X Subset.

In figure 5.7a, the sample time trend forms a half-loop from right to left across the T1-T2 plane, with compact groups of the samples 1 to 57 at the rightmost end of T1 and 144 to 198 at the other end. At around the sample 57, several process changes occurred, i.e., a crude diet change, and the crude unit was known to upset the FCCU several times. The FCCU was running at a high feed temperature from about 1 to 80, and at a low feed temperature from about sample 130 onwards.

The second PC separated samples 203 to 268 from the rest. Operational changes made to handle the contaminated catalyst may have been the reason these samples were separated from the rest.

The plane formed by T1 and T3 provided a good separation of samples 1 to 62 and 80 to 107 from the rest of the data. This latter change appeared to be strongly correlated to the crude diet of the refinery. The crude diet made big swings from an initial light diet to a heavy one at samples 51 and then back to light crude at sample 85. A smaller swing to a heavier diet also took place at sample 107. The remainder of the samples sat left of center in the plane and travelled in a clock-wise fashion with time, probably representing operational action taken to minimize the contaminated catalyst problem on yields and the process.

The remainder of the T plots did not separate out distinct large groups, only small portions of the data, as exemplified by the T1-T4 plane.

This further analysis revealed three operating regions within the subset, these being (approximately) 1 to 62, 80 to 107, and 203 to 268 which appeared to be confounded with crude diet changes. Such a relationship was not evident when the large data set was analyzed and suggests positive benefits in looking at smaller sections of the data.

5.2.3.2 PCA of the Y Subset

Table 5.12 summarizes the statistical results from the analysis. The first three PCs described 62.4% of the cross-validated SS in Y; later PCs had relatively low modelling power and did not reveal further subgroups in the sample space.

Table 5.12–Statistical Results from PCA of Y Subset (Y=268 x 11)

LV	Ordinary %SS Y	Ordinary Cumul. %SS Y	Cross-validated %SS Y	Cross-validated Cumul. %SS Y	% Diff SS Y	Overall CSV/SD
1	41.8	41.8	27.4	27.4	52.6	0.851
2	29.5	71.3	24.4	51.8	20.9	0.761
3	14.5	85.8	10.6	62.4	36.8	0.792
4	5.3	91.1	2.7	65.1	96.3	0.897
5	3.8	94.9	2.6	67.7	46.1	0.841
6	2.4	97.3	1.3	69.0	84.6	0.857
7	1.8	99.1	1.2	70.2	50.0	0.732
8	0.8	99.9	0.6	70.8	33.3	0.512
9	0.02	99.9	0.02	70.8	0.0	0.678
10	0.007	99.9	0.006	70.8	16.7	0.459
11	0.0	99.9	0.0	70.8	n/a	0.542

The T1-T2 plane of Y subset clearly revealed two distinct groupings in the samples; 1 to 62 and the remainder of the data. The accompanying P1-P2 plane suggested that this separation may be related to a distinction between the desirable products propane, butane and gasoline (2,3, and 4) from the undesirable dry gas, heavy cycle oil, and coke (1,7,8 and 11). This change followed a significant shift to heavy crude feed as revealed by the PC analysis of the X space. This sequence of samples showed the plant moving towards production of undesirables.

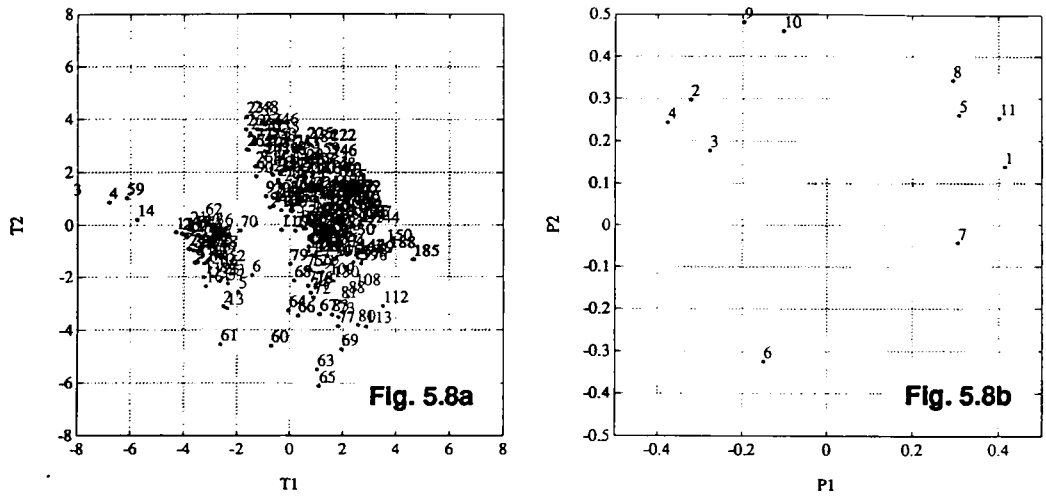


Figure 5.8: Sample and process variable planes from PCA of Y Subset; a) T1-T2 (sample) plane, b) P1-P2 (process variable) plane.

5.2.3.3 Summary of PCA of the Subset of FCCU Data

Examining a subset of initial PCA data revealed further distinct groupings in both the variable space X and the Y product space. In particular, a change in product yields appeared to coincide with the on-set of known feed quality change at the crude unit which was later followed by catalyst contamination and a shift away from production of desirable products to undesirables.

The X space showed three distinct operating regions, which roughly coincided with crude diet changes. As the number of operating points in the data set was reduced

(by examining only a fraction of the original data set) PCA was able to model subtler, and more interesting aspects of the data. With more timely analysis of the data, clearer interpretations of the groupings and changes in the plant should be possible.

CHAPTER 6: PREDICTIVE MODELS AND INTERPRETATION USING PLS

This chapter contains the results of the PLS analysis; building a predictive model of the Y space (product yields and selectivities), further analysis of the data from the PLS plots, and a brief comparison of the results with those of PCA. It also looks at the appropriateness of auto-scaling used in these analyses.

6.1 Development of Predictive Models Using PLS

The primary goal of the PLS analysis was to build a predictive model of the Y space. Since the process data used did not come from a designed experiment the set had to cover as wide a range of operations as would likely be encountered when the model is later used. A secondary goal of the analysis was to gain further insight into the process and the relationship between the X and Y spaces in addition to what was revealed by PCA. Similarities and differences between the PCA and PLS results were also examined.

Tables 6.1 and 6.2 contain the statistical results from the PLS analyses of 136 x variables, 11 y variables and 1170 samples (hourly averages).

Table 6.1--Statistical Results from PLS of X and Y (X=1170 x 136) (Y=1170 x 11)

LV	Ordinary % SS X	Cumul. % SS X	% SS Y	Cumul. % SS Y	Cross-validated % SS Y	Cumul. % SS Y	% Diff SS Y	Overall CSV/SD
1	26.5	26.5	30.4	30.4	30.3	30.3	0.3	0.835
2	13.0	39.5	14.3	44.7	14.2	44.5	0.7	.892
3	13.4	52.9	8.7	53.4	8.7	53.2	0.0	.918
4	5.6	58.5	8.7	62.1	8.5	61.7	2.4	.904
5	4.2	62.7	6.7	68.8	6.4	68.1	4.7	.911
6	4.0	66.7	4.5	73.3	4.4	72.5	2.3	.927
7	3.9	70.6	2.5	75.8	2.5	75.0	0.0	.953
8	1.6	72.2	3.5	79.3	3.2	78.2	9.4	.931
9	2.8	75.0	1.2	80.5	1.2	79.4	0.0	.971
10	1.8	76.8	1.1	81.6	1.0	80.4	10.0	.973
11	1.5	78.3	1.1	82.7	0.9	81.3	22.2	.975
12	1.1	79.4	1.2	83.9	0.7	82.0	71.4	.978
13	1.5	80.9	0.7	84.6	0.5	82.5	40.0	.983
14	1.0	81.9	0.9	85.5	0.7	83.2	28.6	.977
15	0.9	82.8	0.7	86.2	0.4	83.6	75.0	0.987

Table 6.2--Y Product CSV/SD Values from PLS of X and Y

LV	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11
1	0.84	0.99	0.97	0.53	0.68	0.98	1.00	0.66	0.96	0.76	0.60
2	0.62	1.00	0.64	0.99	1.00	0.97	0.80	0.95	0.95	1.00	0.98
3	0.92	1.00	0.98	0.89	0.53	0.94	0.99	1.00	0.89	0.94	0.76
4	0.97	0.76	0.67	0.99	1.00	0.87	0.97	1.00	0.96	1.00	0.97
5	1.00	0.84	0.99	0.90	0.90	0.76	1.00	0.91	0.92	0.94	1.00
6	0.99	0.95	0.82	0.77	0.92	1.00	0.99	0.79	0.98	0.90	0.82
7	0.72	0.97	0.98	0.81	0.86	0.92	1.00	1.00	1.00	0.99	0.87
8	1.00	0.89	0.95	0.90	0.98	0.93	1.00	0.87	0.88	0.91	0.95
9	0.95	0.96	1.00	1.00	0.98	0.98	0.96	0.86	1.00	1.00	0.95
10	0.75	0.99	0.99	0.97	1.00	0.98	1.00	0.97	0.96	0.97	1.00
11	1.00	1.00	1.00	0.97	1.00	0.98	0.98	0.96	0.96	0.94	0.96
12	0.99	1.00	1.00	0.98	1.00	0.99	0.94	0.95	1.00	1.00	0.96
13	1.00	0.97	1.00	0.98	0.98	0.99	0.99	0.83	1.00	1.00	0.99
14	0.92	0.94	0.96	0.98	0.99	1.00	1.00	0.98	0.95	0.98	1.00
15	0.99	0.97	0.99	0.95	0.98	0.98	1.00	1.00	0.99	0.98	1.00

The overall CSV/SD values for the first fifteen dimensions were all less than 1.0 so that all LVs were acceptable for use in the model. The difference between cross-validated and ordinary SS explained per dimension was also very small suggesting that the amount of noise being modeled was relatively small.

For the first nine latent variables, the ordinary percentage sum of squares (SS) explained in Y was very close to the cross-validated percentage SS (see the column "% Diff SS Y" in table 6.1) indicating the strong predictive power of these vectors. The amount of cross-validated percentage SS in Y explained by individual LVs dropped to the 1% level by the tenth latent variable, whereas for the X space it did not reach this level until the fourteenth latent variable. Remember, though, that the ordinary percentage SS in X value would be somewhat inflated compared to the statistic calculated under cross-validation.

Individual CSV/SD values for each y product are also listed in table 6.2. For each vector, the smaller the value is, the greater the amount of variance of that particular y was being explained. For example, the first latent variable explained a substantial amount of variation in the gasoline, LGO and coke yields and selectivities (4,5,8,10 and 11) and to a lesser extent dry gas (1). For the eleventh vector, though, the CSV/SD ratios were no less than 0.94, showing the smaller modeling power of this dimension compared to the earlier ones.

Without looking further to the model plots, determining an appropriate number of vectors for use in the Y space model was not clear-cut from the above information. However, since the emphasis for this PLS model was on the Y space, an initial cut-off was selected at latent variable eleven, making the output space fully dimensional (i.e., there were eleven different yield and selectivity values in the Y space). From these latent variables, a biased regression model for the Y space was built and tested for its fit.

The correlational coefficient R^2, and 95% confidence interval (CI) calculations for this model were compared to two extra cases; using fewer LVs (the first seven) to calculate the regression equations, and using fifteen LVs for the regression development. The results are presented in table 6.3.

Table 6.3--Biased Regression Model Statistics from PLS Analysis

			Model 1 (LVs 1-7)		Model 2 (LVs 1-11)		Model 3 (LVs 1-15)	
Y	Yave	Y Range	R^2	95% CI	R^2	95% CI	R^2	95%CI
1	337.0	241 - 420	0.888	22.6	0.944	16.0	0.956	14.2
2	8.91	4.78 - 11.2	.671	0.824	.773	0.684	.833	0.587
3	11.8	7.84 - 15.3	.896	0.858	.908	0.808	.918	0.763
4	48.1	35.4 - 59.2	.935	2.71	.956	2.24	.967	1.93
5	26.0	17.3 - 34.8	.934	2.53	.941	2.39	.949	2.22
6	2.69	0.0 - 5.51	.709	0.860	.778	0.751	.802	0.710
7	9.77	.662 - 15.1	.420	2.13	.493	1.99	.618	1.73
8	6.37	5.91 - 6.96	.798	0.171	.904	0.118	.946	0.088
9	107.0	93.0 - 117	.530	4.37	.708	3.44	.759	3.13
10	78.1	58.9 - 91.0	.657	5.46	.776	4.40	.810	4.05
11	0.104	.085 - 0.13	0.904	0.0063	0.930	0.0054	0.938	0.0051
			SO: 0.5581 DF(inside):126.1 DF (outside):127 CF: 1.0034		SO: 0.4875 DF(inside):121.7 DF(outside):123 CF: 1.0052		SO:0.4419 DF(inside):117.3 DF(outside):119 CF: 1.0069	

Notes: Yave = Average Y value
SO = standard deviation of X residuals
DF(inside)= degrees of freedom inside training set
DF(outside)= degrees of freedom in test data set
CF = correction factor for training set

Due to the large number of samples, the values of R^2 for all the Y products and all three models were significant at the 95% confidence level. The confidence intervals for

the eleven LV model were from 9 - 14% of the respective y variable ranges, indicating reasonable fits. For most y variables, the fit of the models improved as the number of latent variables used increased. The relative magnitude of the statistics for the individual models also allowed one to rank each y regression equation on the basis of fit from best to worst.

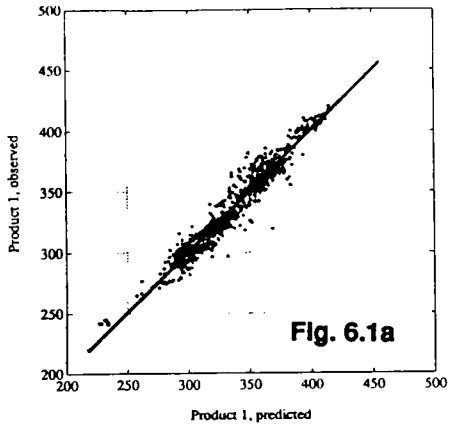
Table 6.4--Relative Order of Biased Regression Model Fits

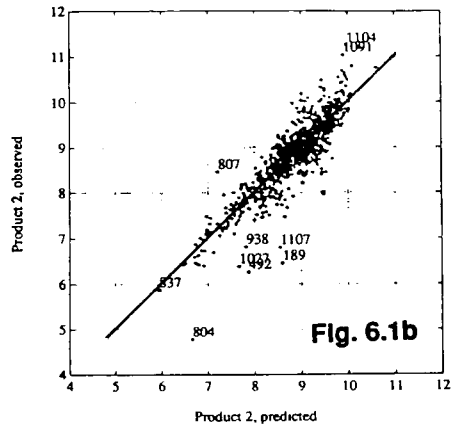
Model 1	Model 2	Model 3	
4	4	4	best fit Y
5	1	1	
11	5	5	
3	11	8	
1	3	11	
8	8	3	
6	6	2	
2	10	10	
10	2	6	
9	9	9	
7	7	7	worst fit Y

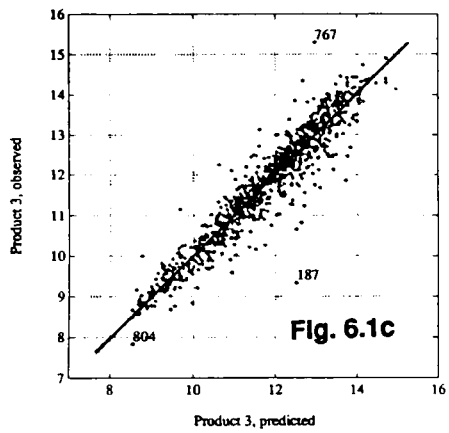
Table 6.4 shows that the relative fit of the Y variables changed somewhat depending upon the number of LVs used for the model. For instance, the 95% confidence interval for gasoline yield (4) dropped from 2.71 volumetric yield percent when seven vectors were used to 1.93 volumetric yield percent when fifteen vectors were used (this represents a 40% reduction in the size of the confidence interval). The same changes in the confidence interval for heavy cycle oil (7) yielded only a 23% reduction, indicating a much smaller improvement with the extra vectors. The rankings, in general, only changed slightly as the number of vectors used for modeling increased.

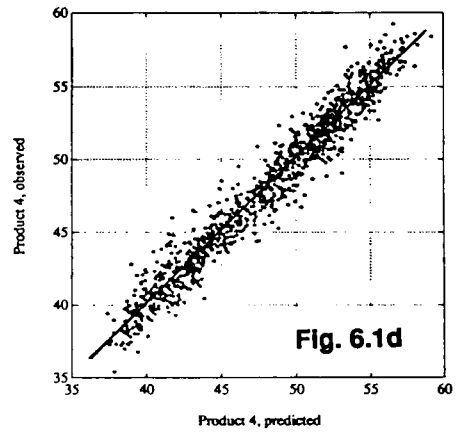
Figures 6.1a-k, which show the plots of the observed versus the predicted y values for all eleven products, revealed the same general trend. Product 4 predictions fit very tightly to the best fit regression line over the entire range covered, while the predictions for product 7 were quite poor. This may be due in part to the fact that the range spanned by product 7 was relatively small yielding little distinctive directionality which is needed for regression modeling.

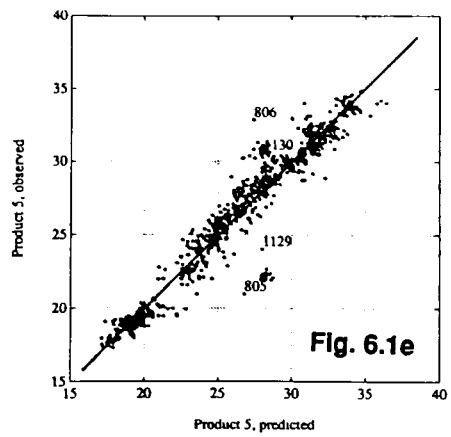
Since the data used to build these regression models was passively collected (i.e., the process was not perturbed in a designed fashion) there are certain limitations on the use of the regression expressions. First, the models are only valid as long as the process remains physically the same and the control and operational strategies used do not change radically. Any of these changes would disrupt the correlational structure amongst the variables and thus make the model invalid. Secondly, extrapolation of linear passive models beyond the range spanned by their reference data set is dangerous as the variance (and hence, the error) of the predicted responses increases monotonically as one moves away from the centre of the reference set (Draper and Smith 1981).

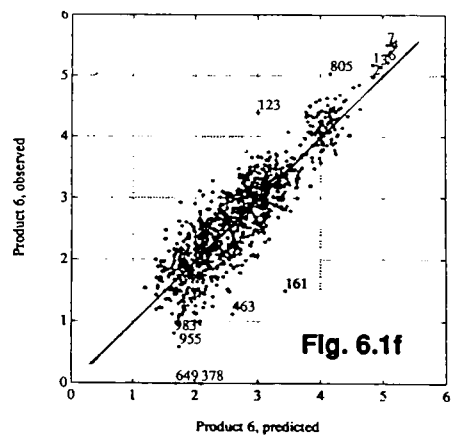












Figures 6.1a-f: PLS predicted Y values versus observed Y values.

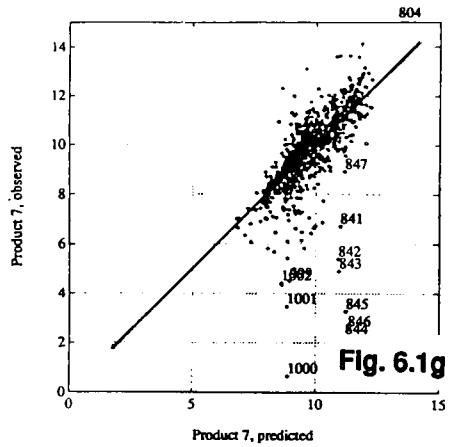


Fig. 6.1g

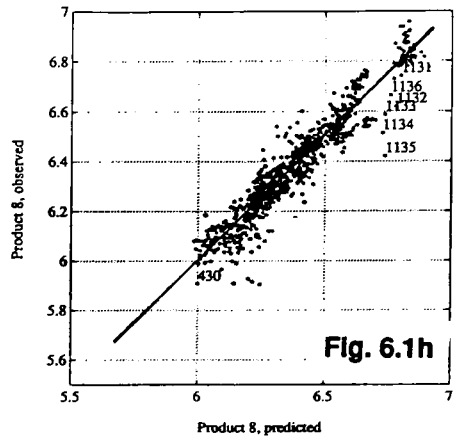


Fig. 6.1h

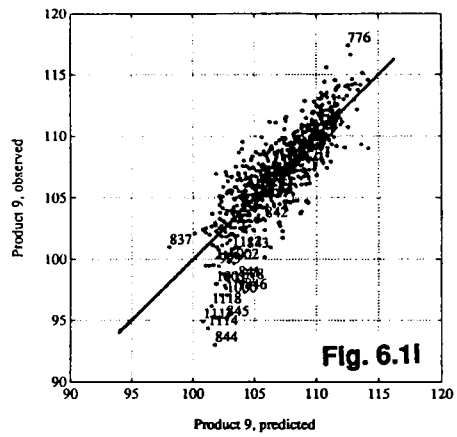


Fig. 6.1l

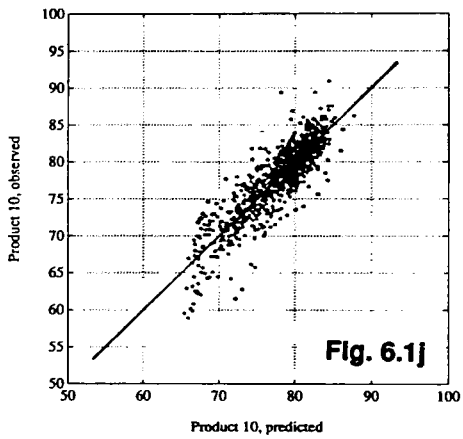


Fig. 6.1j

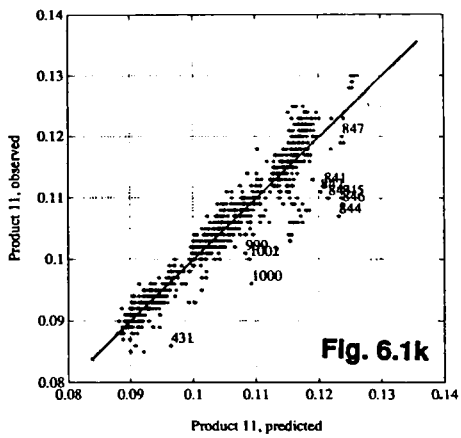


Fig. 6.1k

Figures 6.1g-k: PLS predicted Y values versus observed Y values.

6.2 Analysis of PLS Plots

Table 6.5 lists the findings from examining the T plots (representing the sample or X space), Q plots (representing the product Y relationships) and inner relationship plots of T versus U (representing the X and Y spaces, respectively). These plots are presented in figures 6.2a-k, 6.3a-k and 6.4a-l.

Table 6.5–Analysis of Plots Resulting from PLS Analysis of X and Y

X Space	Y Space	Inner Relationship Between X and Y
T1: Strongly confounded with time trend, change in feed rate, and a known drift in operations. 1-762: High feed rate. 1112-1170: Low feed rate.	**Q1:** Differentiates between gasoline yield and its selectivity (4 and 10) at high feed rates and LGO (5), and coke (8 and 11) at low feed rates. This vector is similar to the PCA of Y results but with clearer separation amongst product groups.	**T1 versus U1:** Many clusters span the full range of these vectors, suggesting that this dimension models strong and contrasting phenomena.
T2: Separates samples based on metal content of feed. 1-7, 398-427,767-823: Low metal feed, and boomer (asphalt) operation versus 550-560: High Ni and V, 824-860: Moderate Ni and high V	**Q2:** Separates the low quality undesirables fuel gas and C.O. (1 and 7) from butane yield (3). Low quality products predominate when feed metal content is high.	**T2 versus U2:** Fairly linear relationship between the X and Y space as exemplified by reasonable fit of samples to the straight line slope.

Table 6.5--Continued

X Space	Y Space	Inner Relationship Between X and Y
T3: Modeling a slow but major shift in the plant, first moving negatively along vector from a score of +8.0 down to -10.0 then returning and surpassing its original position. Interpretation is difficult to make although the Q plot suggests focus is on the top part of the fractionator.	**Q3:** Yields with the largest loadings are gasoline (4), LGO (5), IGO (6), liquid (9) and coke selectivity (11). This suggests that vector is describing partially the top portion of the fractionator (where the 4/5 split is made and 5 and 6 are withdrawn) all of which influence the yield 9. The individual CSV/SD value for 6 is greater than for 4 (therefore less variance in 6 is being explained); its location in the plot suggests the SS modelled in 6 is probably noise.	**T3 versus U3:** Fairly linear relationship; also see the far right cluster in the T space (samples 1112-1170) split into two distinct groups (1112-1129, and 1130-1170) by the vector U. Suggests that a change in the Y space may have occurred.
T4: This LV is modeling samples approaching a unit shutdown and the start-up afterwards. 767-803: Shutdown 804-840: Start-up These points are separated from the rest of the samples.	**Q4:** Propane and butane yields (2 and 3) are modeled in the positive direction while IGO (6) dominates the negative direction coinciding with the start-up period.	**T4 versus U4:** Reasonably linear relationship between X and Y with the start-up data dominating at the lower end of the plot.
T5: Samples are clustering around the zero point of the axis. Since there are no clear outliers, the vector may be describing some overall variability common to all samples.	**Q5:** IGO (6) and proane (2) yields have the strongest influence on the Q vector, followed by products gasoline (4), LGO (5) and coke yields (8) (based on their CSV/SD values).	**T5 versus U5:** Relationship between X and Y is not spanning as large a range as in previous vectors.
T6: Much like T5 in that no cluster is being distinguished from the rest.	**Q6:** Vector separates the desirable product gasoline (4 and 10) from undesirable coke (8 and 11) and from butane yield stock (3). This emphasizes the split between 3 and 4.	**T6 versus U6:** Much like the above plots; little interpretable information.

Table 6.5--Continued

X Space	Y Space	Inner Relationship Between X and Y
T7: Sample group 937-1112 previously unmodelled is split into two groups: 937-1066, and 1068- 1112. They represent two different operating points with a five day gap between them. A similar split was seen in the PCA case and it may be due to a decrease in total feed rate.	**Q7:** Fuel gas yield (1) is separated from the rest.	**T7 versus U7:** Not much interpretable information available.
T8: Much like T5 and T6; tighter clustering of the samples.	**Q8:** Propane (2), coke (8), liquid (9) yields and gasoline selectivity (10) separated from IGO (6). Note that 2, 9 and 10 are coupled through selectivity and yield calculations).	**T8 versus U8:** Range of relationship is narrowing further; only a few points are influencing the direction of the vector.
T9: 804-805: Start-up Dynamic start-up samples are separated from the rest.	**Q9:** C.O. (7) and coke(8) yields, to some extent, dominate the vector.	**T9 versus U9:** A handful of samples are playing a strong role in determining the vector fit: 804-805, 841-846, and 1000-1001. The last two sub-groups are buried in the T plane.
T10: Like T5,T6 and T8, the samples cluster tightly around the zero mark of the vector.	**Q10:** Fuel gas yield (1) is strongly separated from the rest.	**T10 versus U10:** More of the Y space is being modeled, as shown by the wider range covered by the U axis.
T11: Like T10.	**Q11:** Distinguishing which product(s) dominate the LV is becoming difficult; separation amongst Y variables is weaker.	**T11 versus U11:** Can see points that are influencing modeling of the Y space. Not much of the X space is being described by the vector.
T12-15: It appears that these LVs are fitting variability common to most samples.	**Q12-Q15:** The products tend to cluster around the zero of the axes.	**T12-U12 to T15-U5:** As the number of vectors increases, the amount of either space being explained drops off and the linear model becomes less powerful.

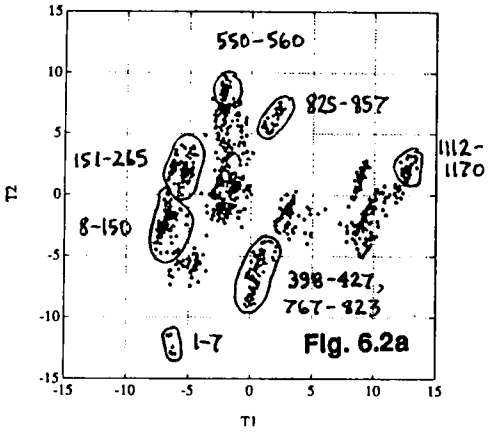


Figure 6.2a: T (sample) planes from PLS.

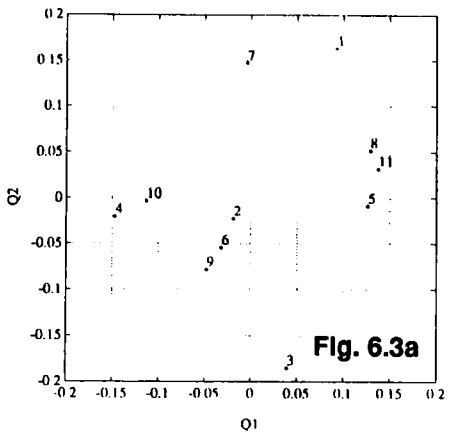
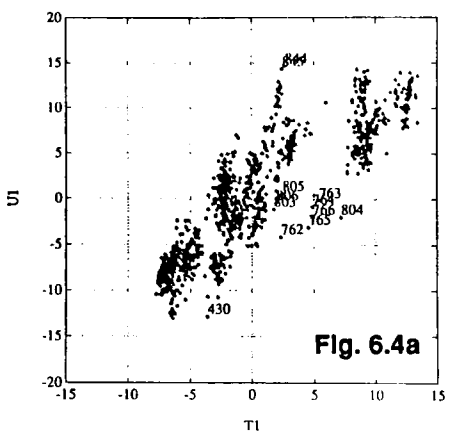
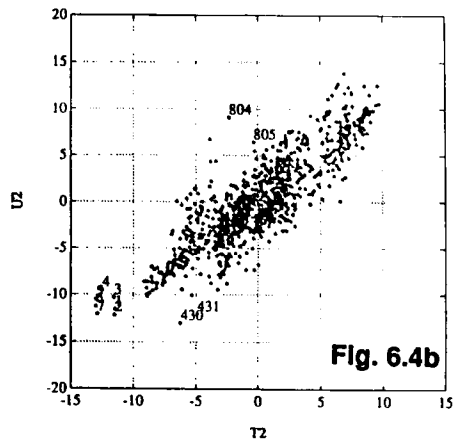
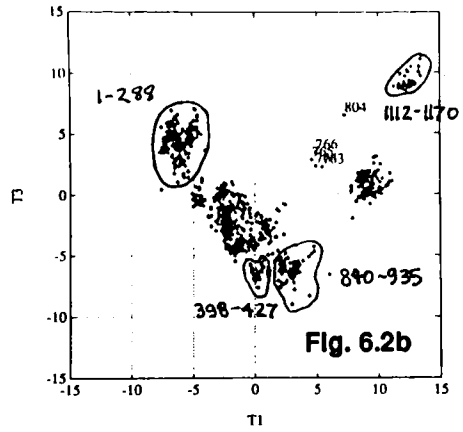


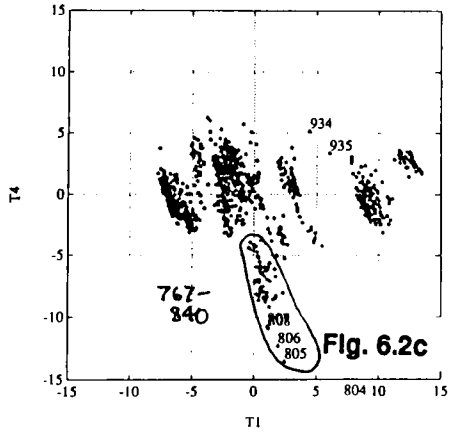
Figure 6.3a: Q (product variable) planes from PLS



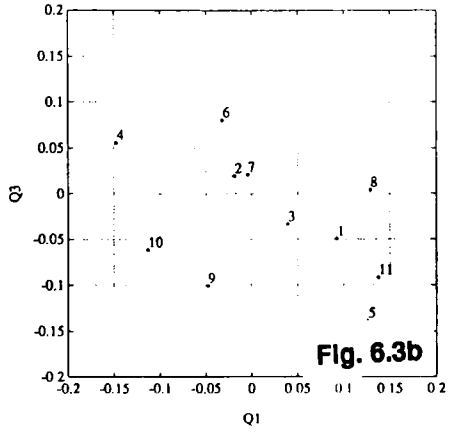


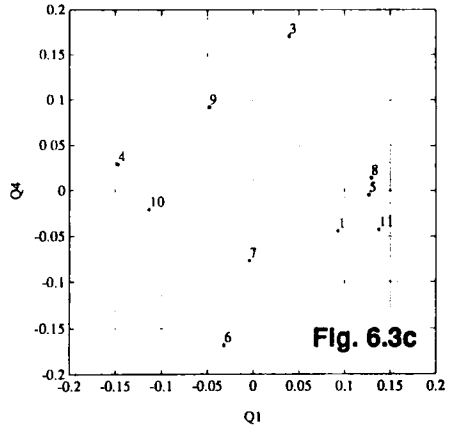
Figures 6.4a-b: T-U inner relationship planes from PLS.



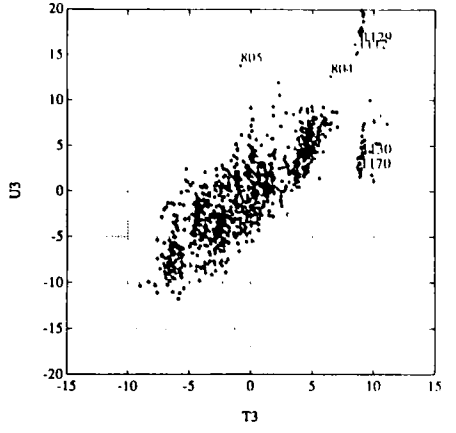


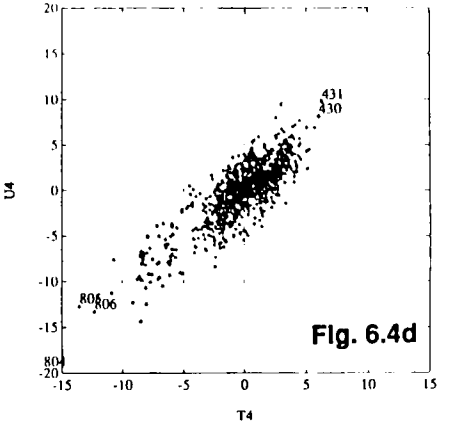
Figures 6.2b-c: T (sample) planes from PLS.



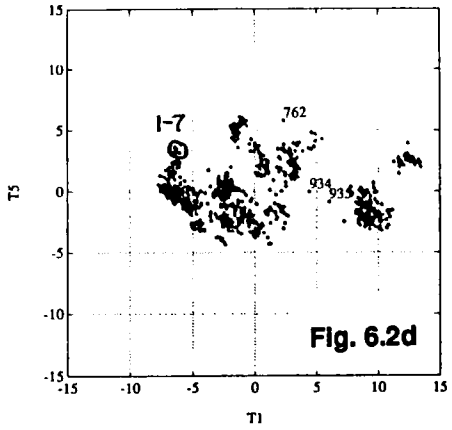


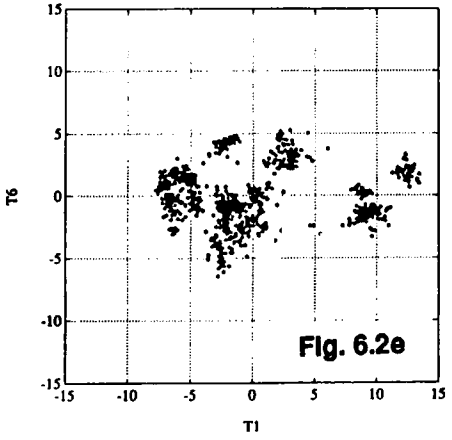
Figures 6.3b-c: Q (product variable) planes from PLS



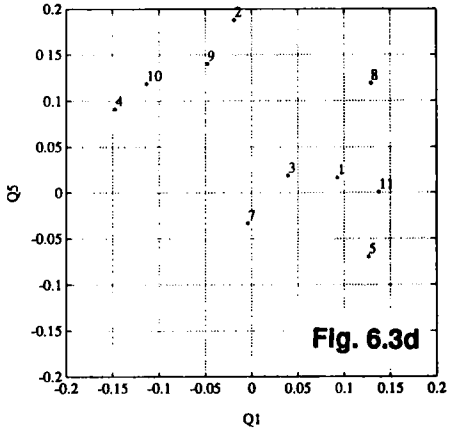


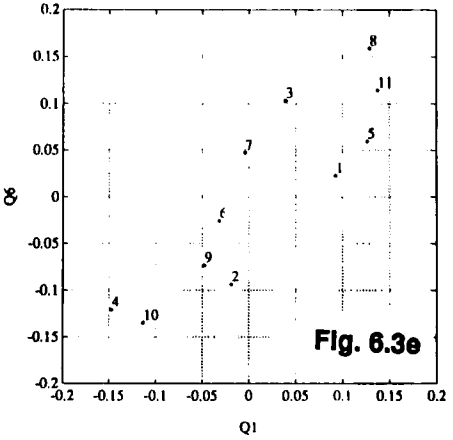
Figures 6.4c-d: T-U inner relationship planes from PLS.



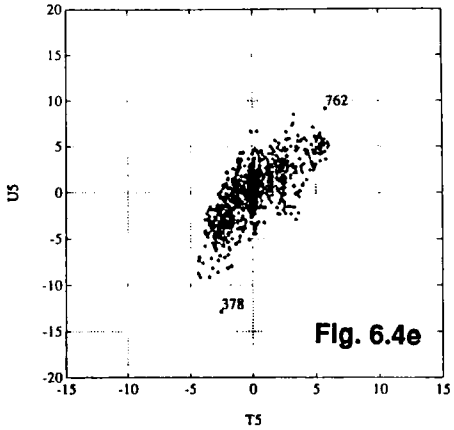


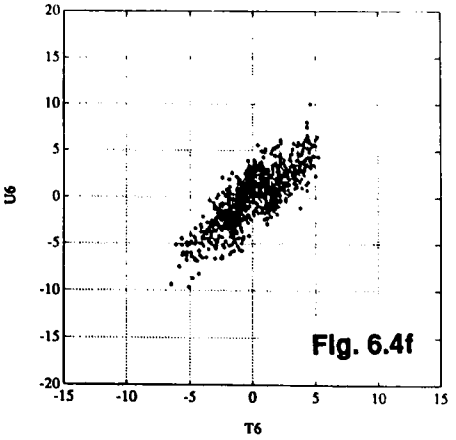
Figures 6.2d-e: T (sample) planes from PLS.



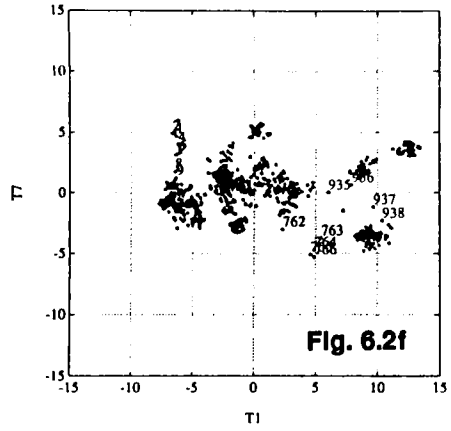


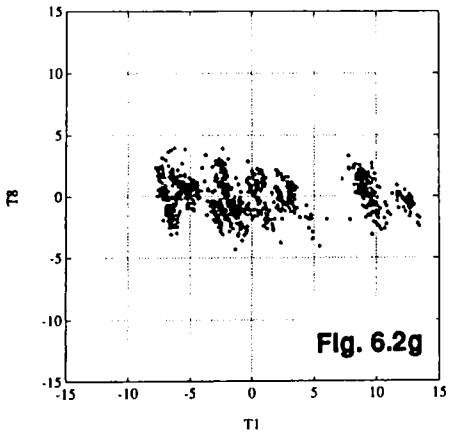
Figures 6.3d-e: Q (product variable) planes from PLS



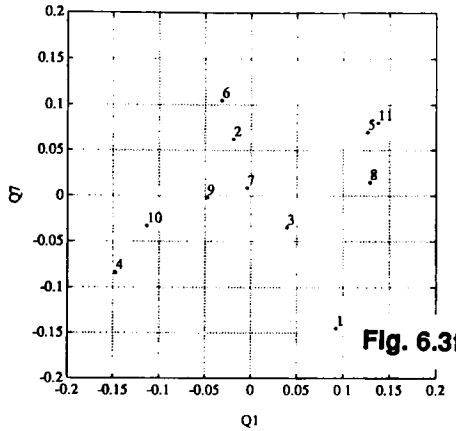


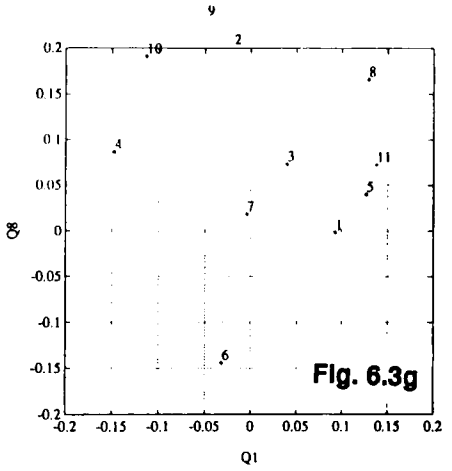
Figures 6.4e-f: T-U inner relationship planes from PLS.



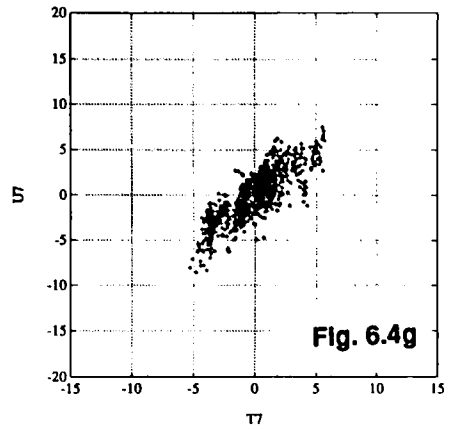


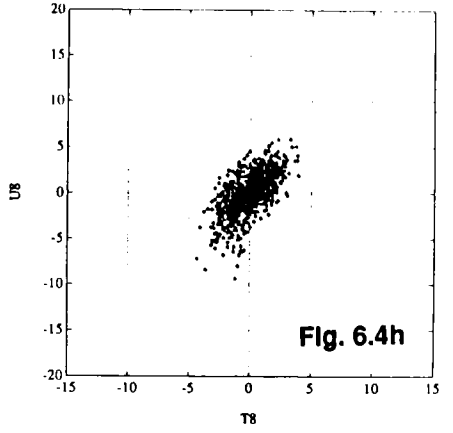
Figures 6.2f-g: T (sample) planes from PLS.



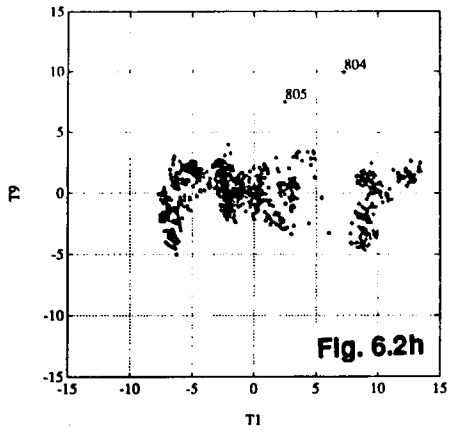


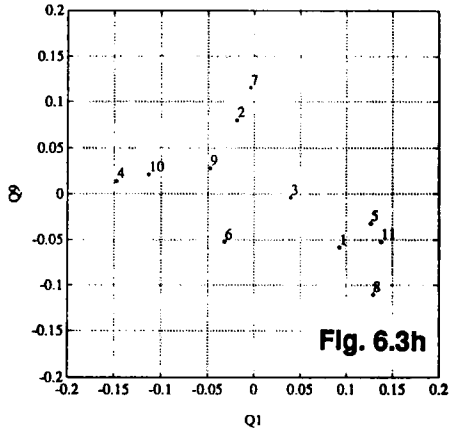
Figures 6.3f-g: Q (product variable) planes from PLS

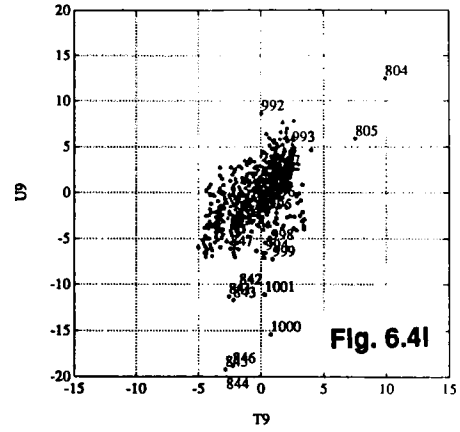


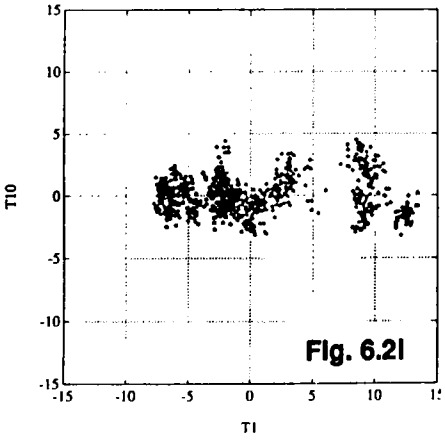


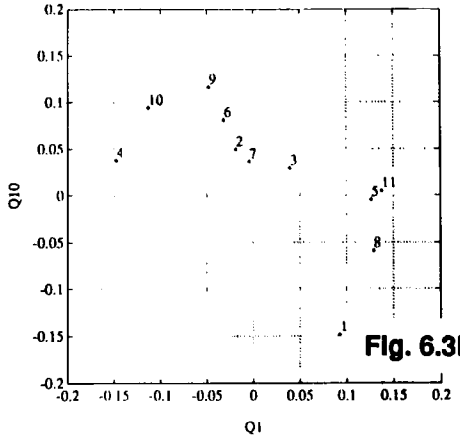
Figures 6.4g-h: T-U inner relationship planes from PLS.

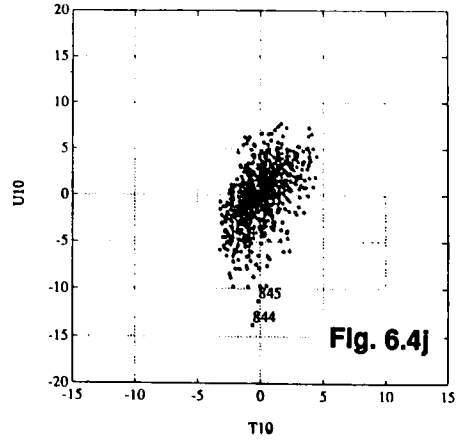








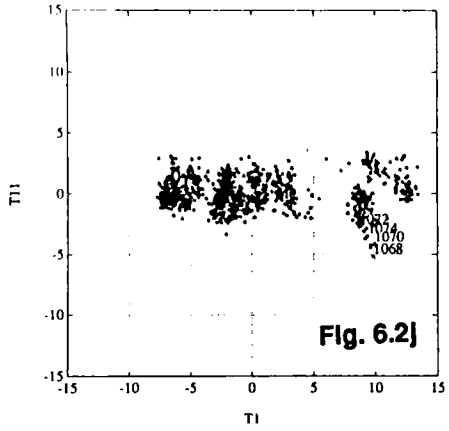


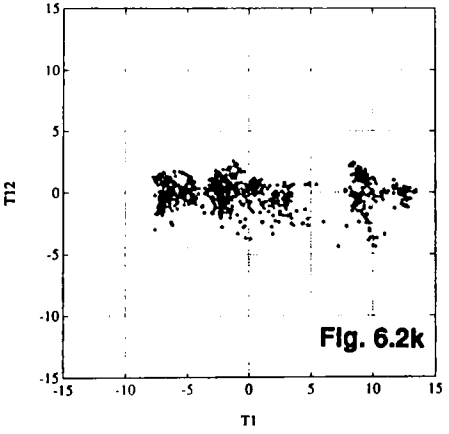


Figures 6.2h-l: T (sample) planes from PLS.

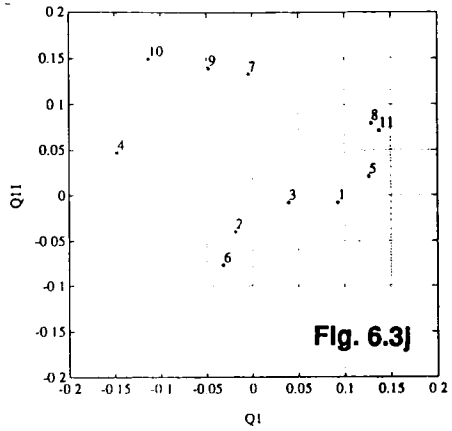
Figures 6.3h-l: Q (product variable) planes from PLS

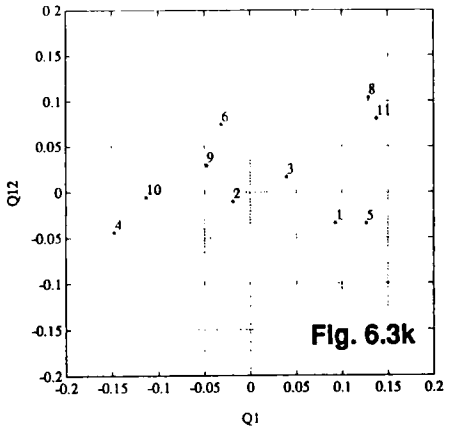
Figures 6.4l-j: T-U Inner relationship planes from PLS.



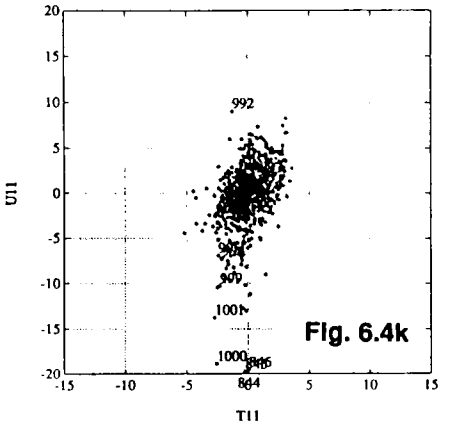


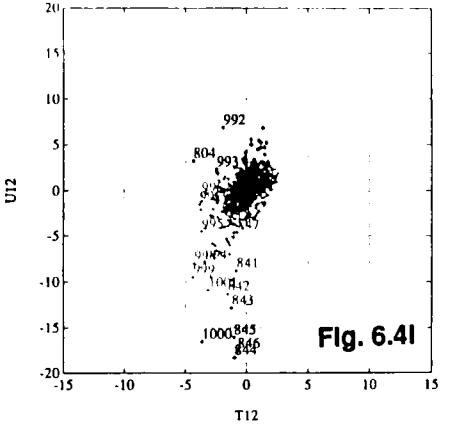
Figures 6.2j-k: T (sample) planes from PLS.





Figures 6.3j-k: Q (product variable) planes from PLS





Figures 6.4k-l: T-U inner relationship planes from PLS.

6.3 Evaluation of the W loadings

In the PLS algorithm, the w weights insure that the vectors t and u will maximize the correlation between the X and Y blocks, whereas the p loadings are calculated so as to keep the t vectors orthogonal. Thus, the w weights provide information about the correlational structure between X and Y, and variables with large weights may be able to give meaning to the latent variables. Due to the sheer number of x variables and vectors used in the model, this evaluation looked at only the first few dimensions. Key process variables (i.e., those with large positive or negative w loadings) were examined to see if changes in their values would correlate to what the vectors appeared to be modelling. A cautionary note: a variable may be highly relevant to the model but due to the fact that it is correlated with many other variables, may have a small loading. Such a variable is hard to distinguish from other variables which have small loadings because they are not highly relevant to the model. This hampers interpreting the LV but actually aids in modelling the data.

By the very nature of the PLS analysis, x variables with large loadings (positive or negative) do have a strong influence on the model. An examination of five of the first six latent variables follows below.

Latent Variable 1:

The Q space differentiated between gasoline mode and light cycle oil (LGO) mode while the T space focused on high versus low feed rate and a known drift in the operations. Some key x variables (with large absolute w loading values) were:

LFD flow from crude unit

Feed rates to both nozzles and the total

Flow of stripping steam

Feed temperature to reactor

Slurry flows at the desuperheater

LGO draw temperature

Temperature of vapor at top of fractionator

The feed rate, feed temperature and stripping steam rate changes reflect the production rate changes, while the draw temperature for LGO (product 5) and end-point temperature for gasoline (4) reflect changes in the product mode.

Latent Variable 2:

This LV distinguished high feed metal content from low; it also separated the lightest and heaviest products (1 and 7) from butane (3). Some key x variables were:

Riser outlet temperature

Regenerator 1 dilute and dense bed temperatures

Regenerator 2 dilute and dense bed temperatures

Catalyst-to-oil ratios

Top temperature and reboiler duty at debutanizer

Feed and reflux flows at depropanizer

Changes in these variables would be expected when moving between light and heavy feeds, and thus, indirectly support the interpretation since higher metals content typically accompanies heavier feed stocks.

Latent Variable 3:

This LV correlated with a slow but major shift in operations over the time span of the data set. The Q space concentrated on the top portion of the fractionator. Some key x variables were:

Dispersion vapor flow

Power to air blower

Air flow rate to regenerator 1

Heat released due to coke combustion

Temperature difference in regenerator 2 gases (dilute - dense)

Upper circulating reflux at fractionator

Reflux rates at top and upper part of fractionator

Temperature of vapor at top of fractionator

Fractionator accumulator pressure

Many of the heavily weighted variables came from the fractionation unit.

Latent Variable 4:

Process operations in this dimension moved towards shutdown and also moved out of start-up phase. Some key x variables were:

Pitch content of crude feed

LFD flow from crude unit

Flow and temperature of LGO reflux at desuperheater

Vapor generation at slurry exchanger

Feed, reflux and reboiler duty of debutanizer and depropanizer

The change in feed rates created by these disturbances would cause changes in some of the key variables listed.

Latent Variable 6:

The Q space separated desirable products from undesirables. Some key x variables were:

Runback flow from crude unit

Dispersion vapor flow

Aeration steam flow

Feed temperature to reactor

Excess oxygen in second regenerator flue gas

Temperature of gas exiting second regenerator

Feed, reflux rate, top and bottom temperature

at debutanizer

Changes in the runback content in the feed and feed temperature support this interpretation; the steam rates could possibly contribute but testing would be needed to verify this.

6.4 Comparing PCA and PLS Results

Since both the PCA and PLS cases described in sections 5.2 and 6.2, respectively, used the same X data set for analysis, it is possible to compare the modeling power and type of information revealed by these two multivariate techniques.

From the statistics tables 5.4 and 6.1, one sees that slightly less of the X space was modelled per PLS dimension than with PCA. This was expected since PLS must compromise by finding a latent variable which explains both the X and Y spaces. The difference, however, in this case was not great. After the first eleven vectors, PCA explained 78.2% of the (ordinary) SS in the X space, while PLS described 76.2% of the X space in addition to 81.3% of the cross-validated Y sum of squares.

The first latent variables extracted by both PCA and PLS analyses appeared to be confounded with the time trend, unit feed rate and production mode changes and general drift in operations. The second and third LVs of PCA appeared to describe feed metals content (or heavier feed in general, the latter LV concentrating on vanadium content) while only the second latent variable from PLS appeared to describe this phenomena.

However, the shapes of the PCA and PLS T1-T3 plane are strikingly similar. If the PCA T1-T3 plot is flipped along the Y axis and tilted slightly to the left, it provides a close fit to the PLS plot, as shown in figure 6.5. This exemplifies the rotation of the X space that takes place as PLS compromises between modeling Y and explaining X. One might then use information from the Q space of PLS to interpret the cluster separations in the PCA space. Thus, the rotation gives the variance in the X space of PLS more meaning.

On the other hand, figure 6.5 suggests that the single third latent variable in PLS describes what the combination of the first and third latent variables in the PCA space appear to model: vanadium metal content of the feed plus some of the confounded process changes dominating the first LV.

For the later dimensions, PCA was free to describe characteristics separating large clusters of points whereas PLS (due to the influence of the Y space) concentrated on modeling the process changes which influence the product outputs. This means that PCA could reveal more subtle changes in the X space, while the advantage of PLS was the availability of the Q space to help explain changes in the T planes. For instance, PCA's latent variable four broke apart groups F (936 to 1065) and G (1066 to 1111) but identifying a plausible cause for this splitting was difficult. In the PLS analysis, however,

the same separation was modelled by the seventh latent variable. Added information from the Q plane suggested this may be due to a product yield shift away from product 1 and towards product 6.

Such an interpretation was checked by studying the yield values for products 1 and 6 during these time intervals. Figures 6.6 and 6.7 show that these two products exhibited distinct changes in yield at the sample interval 1065 to 1066.

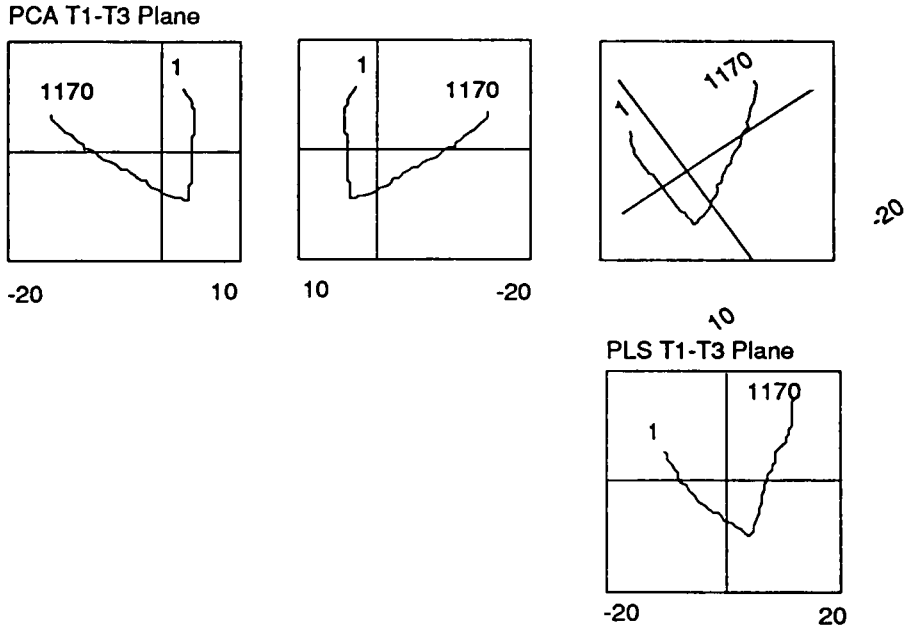


Figure 6.5: Rotation of PCA T1-T3 plane to match PLS T1-T3 plane.

Thus, one can see that PCA and PLS reveal different pieces of information about the process. PCA is free to reveal subtle changes or clusters in the operating data, which may, however, have little relevance to the Y space. On the other hand, PLS must compromise between explaining the X and Y spaces at the same time but the extra information provided by the Y space can aid in the interpretation of the operational shifts. It also does not reveal subtle operational changes as easily (or as early on in the analysis) as PCA.

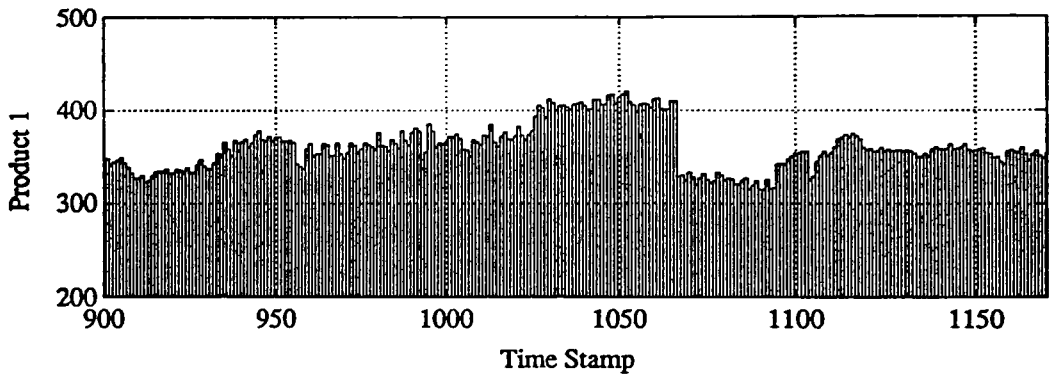


Figure 6.6: Time series plot of product 1.

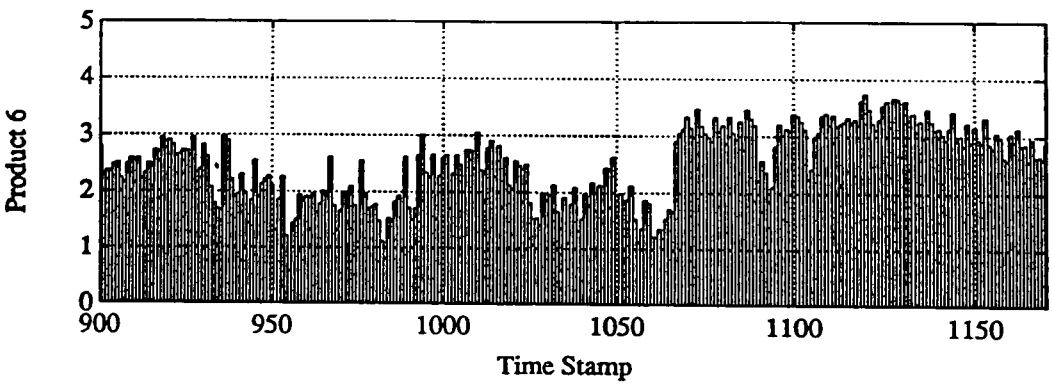


Figure 6.7: Time series plot of product 6.

This brief comparison shows that it is important to have a clear idea as to the purpose of the multivariate analysis before selecting which approach (PCA or PLS, or both) to use.

6.5 Re-Scaled Case

The appropriateness of auto-scaling is a more important issue in PCA than PLS because PCA only has the variance structure of the single block to work with, whereas PLS focuses on maximizing the inner relationship between two blocks.

The main concern in using auto-scaling for these analyses is the scaling up of variance for variables which are almost constant. If these variables dominate the PCA or PLS models and their small variance represents noise, then the model is being dominated by noise. Auto-scaling is not beneficial in this case as it is only inflating the amount of sum of squares (SS) to be explained and probably swamping out more important process information. Such variables should not be scaled at all and the analyses re-run.

If these variables dominate the models, but their variance represents natural process variability, then this information should aid in interpreting the model. The problem then lies in distinguishing between noise variance and normal process variance.

Of course, if these variables do not dominate the PCA or PLS models, then there should not be a problem with using the auto-scaling approach.

Loadings from the three models (PCA on X, PLS predictive model and SPC monitoring reference data) were examined along with their auto-scaled weights to see if variables which were scaled-up due to auto-scaling dominated the latent variables. In the case of PCA, some scaled-up variables contribute heavily to the principal components. This is expected since PCA works solely with the variance structure of the single data block.

In the case of PLS, feed rates (which have a large auto-scaled weight) dominate the first vector. The data set for this analysis spans a large operating region and since the feed rate has a strong influence on unit operations, such domination is expected. Auto-scaled up variables for the other PLS dimensions and the SPC models do not dominate the models. In these cases, the inner relationships between the blocks is important and influences the w loadings and the direction of the latent variables in general.

Thus, by giving all variables equal variance, auto-scaling will probably inflate the noise component of some variables. In these analyses, auto-scaling appears to have a stronger influence on PCA results, due to its focus on the variance structure, and less so on PLS analyses, where the inner relation between two blocks is being maximized. The appropriateness of the type of scaling used is heavily dependent on the data being studied and the analysis to be used, and must be considered in this light. For a first look at the FCCU process data, auto-scaling does not appear to unduly influence the model in a negative manner.

6.6 Summary of Observations from PLS analysis

Despite the undesigned nature of the data set and the number of process changes taking place within it, PLS was able to generate a good predictive model for most of the y variables. The first eleven LVs captured 81.3% of the cross-validated SS of the Y space; this was significantly better than the maximum fit of only 44.2% of the cross-validated SS of Y generated by PCA. It appeared that the "randomness" in the Y space left unmodeled by PCA actually correlated strongly with the operating conditions of the X space, thus yielding a better fit with the PLS analysis.

Cause and effect interpretation of the LVs was much more difficult. Major process changes in production rate and production mode, feed quality, a known drift in the process operations, and large transients were explained in the first six to seven LVs; later dimensions modelled smaller, sporadic process changes.

The similarities in the first few LVs of both PCA and PLS revealed that the dominant variations in the data set were also the dominant operational changes. The differences in modelling objectives was also illustrated by comparison of the T1-T3 planes from both analyses.

The appropriateness of auto-scaling for all analyses (PCA, PLS and SPC) was considered. For a first look at this FCCU, it did not appear to unduly influence the model in a negative manner.

CHAPTER 7: MULTIVARIATE SPC MONITORING SPACE

This chapter investigates the development of multivariate statistical process control (SPC) charts for monitoring the behaviour of the FCCU over time. A monitoring space representing "normal operations" is built using PLS from a select set of samples called the reference data. A second set of samples representing abnormal operations is used to test the resolution of this monitoring space and to show how different combinations of latent variables can be combined to form meaningful monitoring planes. A method for uncovering the possible causes of abnormal process changes is also proposed and discussed.

7.1 Overview

The purpose of developing a monitoring space was to build a low dimensional Shewart-type chart in which the process could be constantly monitored for abnormal behaviour or events. As seen in the last two chapters, the FCCU was not operated at one steady-state condition during the data collection but rather was in a constant state of transition. Therefore, the appropriateness of an SPC monitoring approach would be questionable. However, the approach was developed anyway, and its abilities and limitations discussed.

The process data set offered a wide range of operating "windows" from which to select a "normal" or reference data set. Selection of the final reference set was heavily

dependent upon what events or changes were considered "abnormal". For this analysis, only non-boomer (regular production) data was selected. This should eliminate large feed quality and production mode changes. Samples were restricted to the month of October when plant operations were considered relatively smooth. Data saved for the test set (to later verify the model) included abnormal periods of high and low catalyst activity, high and low nickel (Ni) and vanadium (V) metals content in the FCCU feed, plant moves into and out of boomer (or asphalt, B) and non-boomer (regular, NB) operations and a period where the upstream crude unit created disturbances in the FCCU. Tables 7.1 and 7.2 provide a complete listing of the normal and abnormal data used for the SPC evaluation.

Table 7.1--Multivariate SPC Normal Data Reference Set

Sample Number	Date	Sample Number	Date
1 44	Oct 02 3:00 Oct 03 22:00	144 214	Oct 12 10:00 Oct 15 8:00
	$2\frac{1}{2}$ days later...		5 hours later...
45 136	Oct 06 13:00 Oct 10 9:00	215 257	Oct 15 13:00 Oct 17 7:00
	9 hours later...		$2\frac{1}{2}$ days later...
137 143	Oct 10 18:00 Oct 10 24:00	258 290	Oct 19 16:00 Oct 20 24:00
	$1\frac{1}{2}$ days later...		

Table 7.2--Multivariate SPC Test (Abnormal) Data Set

Date	Event	Sample No.
Sep 21 16:00 Sep 24 8:00	Start-up and move into boomer operation.	1 65
Oct 1 16:00 Oct 2 2:00	Transition into non-boomer mode.	66 76
Oct 3 23:00 Oct 4 8:00	High catalyst MAT (64).	77 86
Oct 11 1:00 Oct 12 9:00	3:00 Product 2 off-specification, too much gas, air blower limited. 20:00 High product 7 yield.	87 119
Oct 21 1:00 Oct 22 3:00	Cooling water temperature trouble.	120 145
Oct 22 4:00 Oct 22 10:00	Transition into boomer mode.	146 152
Oct 26 11:00 Oct 26 20:00	Transition into non-boomer mode.	153 162
Oct 26 21:00 Oct 28 24:00	Normal data, but later than the reference data.	163 212
Oct 29 1:00 Oct 29 11:00	Crude unit upset FCCU several times.	213 223
Oct 29 12:00 Oct 30 24:00	Normal data, but later than the reference data.	224 260
Oct 31 1:00 Nov 4 24:00	High metals (V and Ni) in feed.	261 356
Nov 5 1:00 Nov 7 14:00	Normal data, but later than the reference data.	357 414
Nov 7 16:00 Nov 8 8:00	High catalyst MAT.	415 431
Nov 9 10:00 Nov 12 8:00	Low catalyst MAT.	432 502
Nov 18 22:00 Nov 19 23:00	High vanadium content in feed.	503 528
Nov 20 1:00 Nov 20 8:00	High vanadium content in feed.	529 536
Nov 30 9:00	Normal data, but later than the reference data.	537

If no Y data is available, PCA can be used to form a monitoring space that would detect changes which disturb the relationships amongst the x process variables. However, the variables with the largest variances may not necessarily be the most informative. If Y data is available, it can be used with PLS to create a more knowledgeable monitoring space where changes in the X planes can be related to changes in the Y space. This was the approach selected for the analysis. (If Y data is not available but can be inferred from specific x variables, then one could also build a model using X versus the inferential, now "Y" data.) For this analysis, all eleven measured y variables were used.

The approach to be followed in defining the normal operating region is that of Kresta, MacGregor and Marlin (1991a) where the boundary of acceptable behaviour is determined such that it encompasses a certain percentage of the reference samples. The sum of the prediction errors (SPE), plotted as a function of sample number, act as a guide for detecting gross deviations of the test data from the reference set. Since these values should be random in magnitude and independently distributed, reference samples with abnormally large SPE values should be examined for appropriateness in the building of the monitoring space.

The T plane monitors the structure amongst process variables while the SPE values signal a poor model fit. The individual focus of these two types of plots should help unveil the causes of process deviations. For instance, if a process change causes a large move in one or more of the process variables but which does not change the basic relationship with the Y space, then one will see a shift in the T space without a large SPE value being generated. If, however, the change is caused by an event not captured in the reference data set, then it may show up not only as a shift in the T planes but also generate a significantly large SPE value indicating a change in the Y space (Kresta, MacGregor and Marlin 1991a).

7.1.1 Determining the Dimensionality of the Reference Model

The statistical results for the reference data using PLS are provided in tables 7.3 and 7.4. The statistical significance of the overall CSV/SD values for all fifteen latent variables would allow all of them to be considered for the model. The value for the percentage difference between the ordinary and cross-validated percentage SS in Y, however, suggests that LV twelve and onwards may have been modelling process noise and could be dropped from the model. This is also suggested by the individual CSV/SD values for the y variables in table 7.4, where values near unity for the later dimensions imply that little predictable process variability was modelled.

Table 7.3–Statistical Results from PLS Analysis of Reference Data (X=537 x 136) (Y= 537 x 11)

LV	Ordinary %SS X	%SS Y	Cross-validated %SS Y	Cumul. %SS Y	% Diff %SS Y	Overall CSV/SD
1	23.1	31.0	30.8	30.8	0.6	0.832
2	16.9	31.6	31.2	62.0	1.3	.740
3	12.0	5.6	5.2	67.2	7.7	.927
4	6.2	6.6	5.7	72.9	15.8	.906
5	6.8	3.1	2.3	75.2	32.6	.953
6	4.6	2.5	1.8	77.0	37.2	.959
7	3.6	2.1	1.7	78.7	21.8	.955
8	1.8	3.5	2.8	81.5	25.0	.915
9	1.6	1.9	1.3	82.8	46.2	.954
10	2.6	0.7	0.4	83.2	75.0	.982
11	1.5	0.9	0.5	83.7	80.0	.978
12	1.2	0.7	0.2	83.9	250.0	.988
13	1.0	0.9	0.1	84.0	1451.7	.997
14	1.2	0.5	0.3	84.3	100.0	.986
15	0.7	0.7	0.1	84.4	775.0	0.995

Table 7.4–Y Product CSV/SD Values from PLS of Reference Data

LV	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11
1	0.99	1.00	0.69	0.67	0.28	0.66	1.00	0.98	0.95	0.90	0.71
2	0.37	0.96	0.58	0.67	0.77	0.96	0.65	0.67	0.86	0.92	0.46
3	0.83	1.00	0.96	0.83	0.89	0.86	0.92	0.99	0.94	0.85	0.92
4	0.98	0.87	0.88	0.85	1.00	0.85	0.92	0.87	0.93	0.98	0.98
5	1.00	0.98	1.00	0.90	0.98	0.97	1.00	0.84	0.94	0.94	0.99
6	1.00	0.96	0.98	0.98	0.93	0.98	0.78	0.98	1.01	1.00	0.84
7	0.98	1.00	1.00	0.99	0.98	0.96	0.77	1.00	0.94	0.94	0.94
8	1.00	0.82	1.01	0.80	0.97	0.99	1.00	0.98	0.90	0.89	1.00
9	0.97	0.99	0.96	0.96	1.00	0.99	0.99	0.93	0.89	0.93	1.00
10	0.86	1.01	0.94	1.00	0.95	0.97	1.00	1.00	0.96	0.99	1.00
11	0.99	0.99	1.00	1.00	0.99	1.02	1.00	0.92	0.95	0.97	0.94
12	1.00	0.98	0.99	0.94	0.97	1.00	0.99	1.00	1.00	0.98	0.98
13	0.96	1.00	1.02	0.99	1.00	1.00	1.00	1.00	0.99	0.98	0.99
14	0.87	0.99	1.00	1.00	0.99	0.99	0.95	1.00	1.00	0.99	0.99
15	0.99	1.01	1.02	0.96	0.96	1.00	1.01	0.96	0.99	0.99	1.00

A compromise must be made between low dimensionality and high resolution in selecting the final dimension for the reference model. All dimensions of the reference model must be monitored, and thus, it is best to restrict this number to a reasonable size (appropriate to the specific monitoring case). However, if the reference data is somewhat complex and non-linear, a large number of latent variables may be needed to model it sufficiently. Inadequacies in this balance can be expected to show up as poor resolution of the abnormal samples from the normal operating regions.

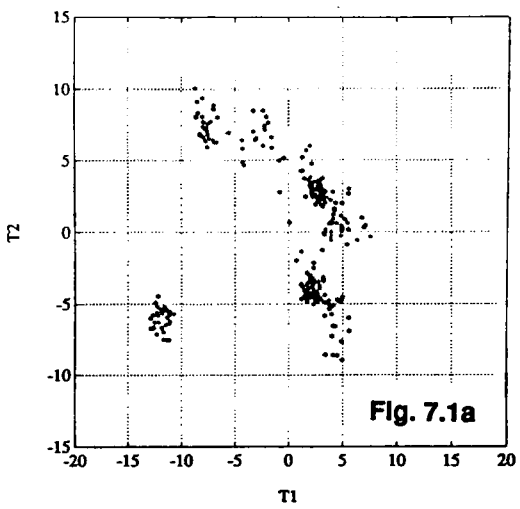
For this study, the first eight LVs were selected for use in the reference model. Pairs of latent variables were plotted against each other (rather than all against the first latent variable) to yield four monitoring planes plus one SPE plot as shown in figures 7.1a-d and 7.3, respectively.

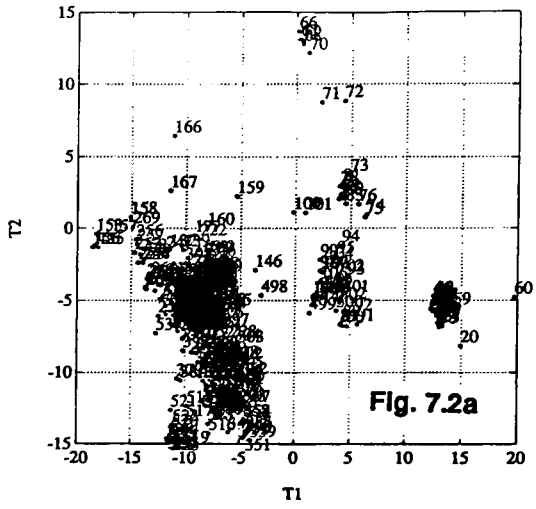
The reference SPE plot, figure 7.3 was generated by fitting the reference data samples to the PLS model using the first eight LVs and summing up the squares of the

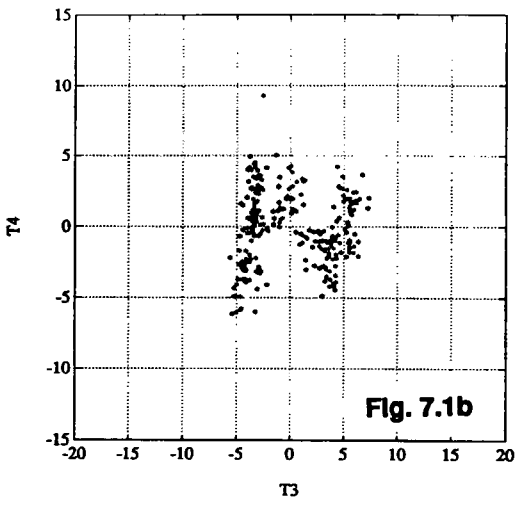
prediction errors for each y variable, per sample. For a 95% confidence level, the cut-off value (so that 5% of the reference SPE values would be greater than the cut-off) would be approximately 50. Test sample SPE values greater than this would be considered abnormal.

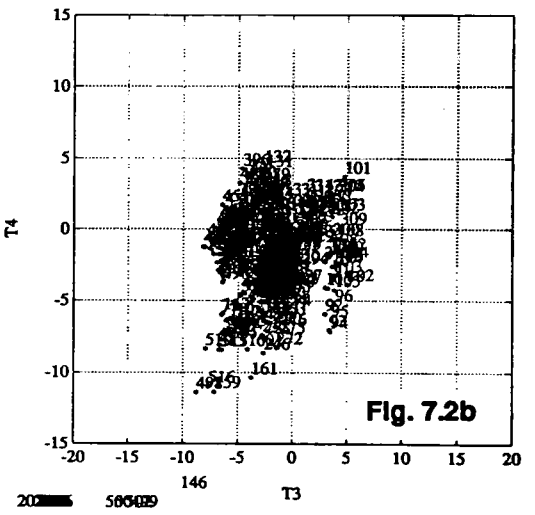
The region spanned by the reference data in the T1-T2 plane was the largest of any monitoring planes created and explained 40% of the ordinary SS in X and 62.0% of the cross-validated SS in Y. It contained two distinct clusters which complicated defining the "normal" region. The small cluster of reference points in the lower left quadrant of figure 7.1a represents samples 258-290 which are two and a half days later than sample 257. One could accept these two clusters as separate yet legitimate regions for operation, but it is more likely that the reference data set did not contain enough samples to properly model the normal region in this part of the plane. Figure 7.3 showed that the SPE values for the left most data group in the reference set were not abnormally larger than those of the earlier data.

The T3-T4 and T5-T6 planes also appeared to contain more than one reference cluster which made determining a boundary for normal operations tenuous. Due to the process changes which are typical of this unit, it is not known whether further samples would aid in filling the gaps between these reference clusters or if more data would merely extend the reference space into a new direction. The reference planes were thus used without boundary limits.



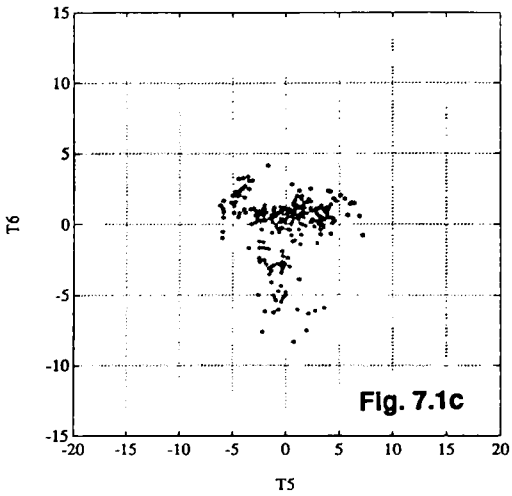


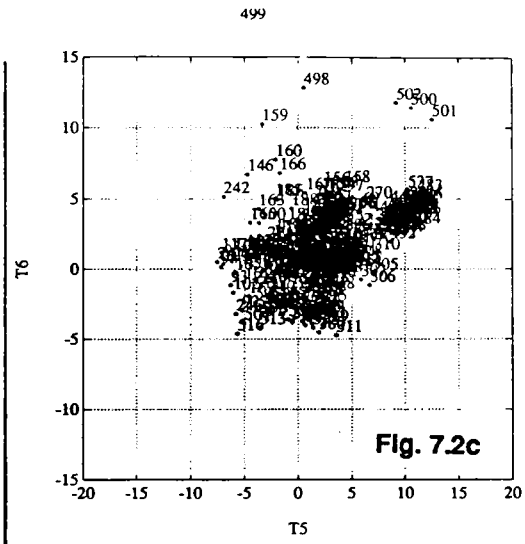


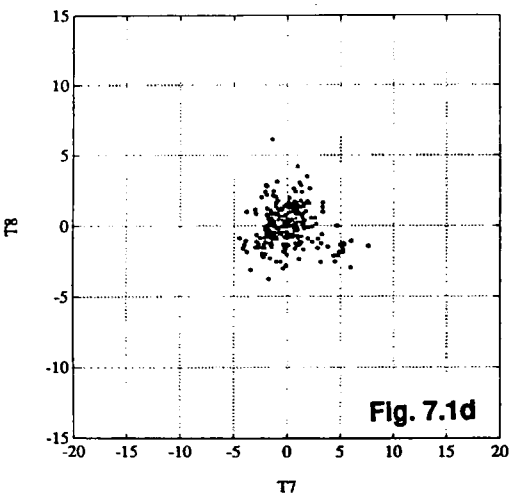


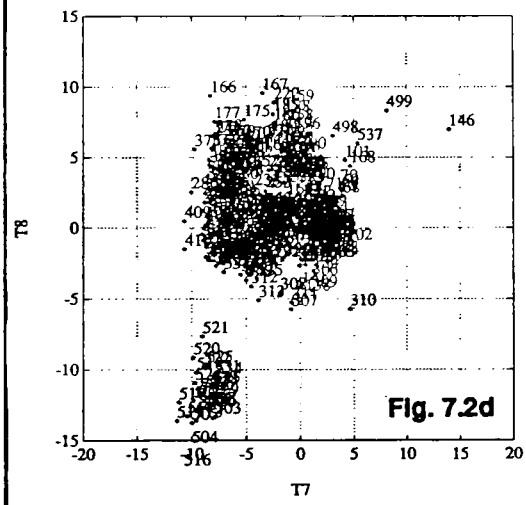
Figures 7.1a-b: Reference SPC T monitoring planes.

Figures 7.2a-b: Test data plotted in SPC monitoring planes.









Figures 7.1c-d: Reference SPC T monitoring planes.

Figures 7.2c-d: Test data plotted in SPC monitoring planes.

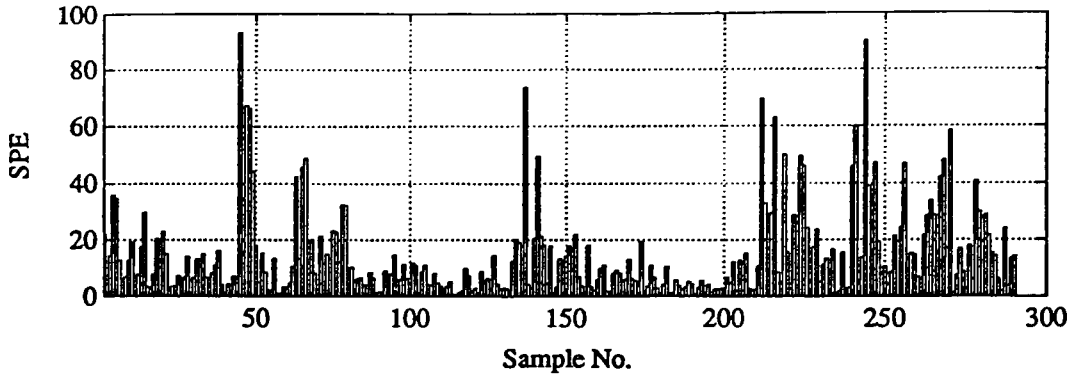


Figure 7.3: Reference data SPE values.

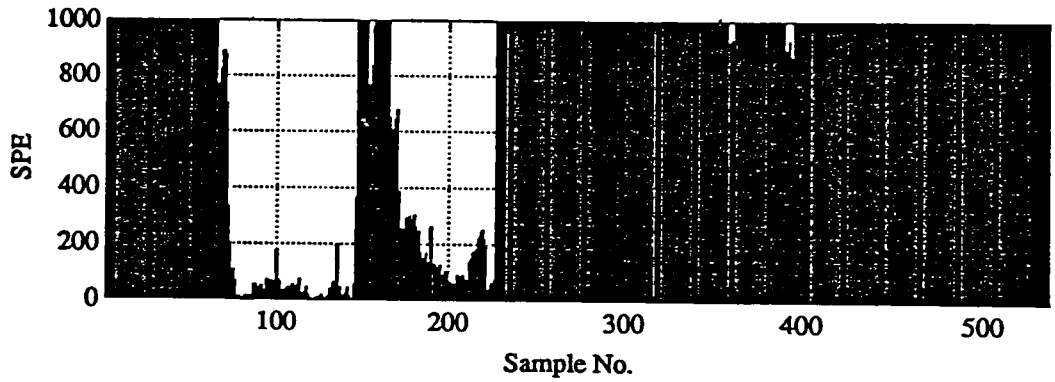


Figure 7.4: Test data SPE values.

7.1.2 Evaluation of the SPC Monitoring Space

The test (abnormal) data was fitted to the SPC reference model and the resulting T scores were plotted in figures 7.2a-d. The T scores for samples 1 to 65

(representing a start-up phase and move into boomer operation) were so large, they lay off the portions of all four planes shown. In the T1-T2 plane, samples 66 to 72 lay directly above the normal operating region. Samples 1 to 76 precede the time period selected for the reference set, and essentially all the test samples following the time period selected for the reference set resided in the lower left quadrant of the plot, suggesting that the first two latent variables were heavily confounded with the time trend observed in the data. Thus, this first plane did not prove to be of much value in terms of monitoring (although the vectors themselves played an important role in the modeling of the reference space).

In the T3-T4 plane, the test data generally lay within the region defined by the reference data, indicating that latent variables 3 and 4 still explained variation common to most samples.

In the T5-T6 plane, samples 432-497 representing low catalyst MAT activity clustered at the coordinates (10,5) which is clearly outside the normal operating region. Samples 498-502 representing preparation for a boomer run and samples 159-160 (part of a transition into a non-boomer operation) lay between +7 and +15 in the T6 direction.

In the T7-T8 plane, feed metals as well as major process moves were distinguished from the normal reference region. Samples 503 to 537, representing high vanadium content in the FCCU feed lay furthest from the normal region, in the lower left quadrant. Although more difficult to see in these plots, samples 261 to 356 representing high vanadium and nickel in the feed clustered in the area of coordinates (-7,-2) to (-3,-2) which just borders on the normal operating region. Samples 357 to 414, which represent normal but later data than the reference set, clustered in the region of (-5,5). Samples 1 to 65, 147 to 152, and 498 to 502 all representing major transitions into or out of boomer operations (or start-ups) had T scores too large to lie in the region of normal operation.

SPE values for the test data were calculated and plotted as a function of sample number in figure 7.4. Clearly, samples 1 to 75 and 145 to 152 do not belong to the reference region, as was also indicated in the T plane plots. Although test samples 163 to 212 represent normal operations, they represent samples approximately six days after the reference data and showed significantly large SPE values compared to the reference SPE value. This showed that perhaps the reference set did not contain a large enough sampling of data from normal operations. As the sample times moved further away from the reference data, SPE values increased accordingly, showing that the process operating conditions were continually shifting. Samples 230 to 498 had SPE values around 2000, while samples after 500 had SPE values of approximately 4000. Continuously high SPE values for "normal" samples might act as a good indicator of when model updating would be needed.

Table 7.5 summarizes the abnormal periods flagged either through their location in the monitoring planes or their SPE values. The periods of start-up and major transitions lay far from the normal operating regions defined in all four planes and were accompanied by extremely large SPE values.

The first four vectors of the monitoring space appeared to describe variation which was common to both normal and abnormal data. This made the plane formed by these vectors a difficult space to monitor and its only use might be to detect radically large changes in the process (however, operators are sure to be aware of such changes).

Subtler changes affecting the unit, such as feed quality and catalyst activity were better distinguished by later vectors. These abnormalities broke out in asymmetrical patterns, i.e., the normal operating region was modeled as a distinctive cluster in the monitoring plane and the abnormalities were found to lie somewhere outside of this region.

Table 7.5--Abnormal Events Flagged by the SPC Monitoring Procedure

Sample	Large SPE Value	T1-T2 Plane	T3-T4 Plane	T5-T6 Plane	T7-T8 Plane	Abnormality
1-65	*	*	*	*	*	Start-up of unit
66-76	*	to 72	*	*	*	Move into NB operation
77-86						High catalyst MAT
87-119	100					Y2 off-specification, etc.
120-145	135	*				Cooling water T trouble
146-152	*	*	*	*	*	Move into B operation
153-162	*	*		159-160		Move into NB operation
163-212	up to 199	*				Later than reference data
213-223	*	*				Crude unit upset FCCU
224-260	228+	*				Later than reference data
261-356	*	*			*	High Ni and V in feed
357-414	*	*			*	Later than reference data
415-431	*	*				High catalyst MAT
432-502	*	*		*		Low catalyst MAT
503-528	*	*			*	High V in feed
529-536	*	*			*	High V in feed
537	*	*			*	High V in feed

Note: * - indicates that all samples in the group were flagged as abnormal.

The planes also suggested the degree of change taking place in the test data; from radical changes evident in the first three dimensions, to subtler changes in the process only revealed by the later latent variables.

The fact that some abnormal events were not well distinguished from the reference model may be due to these reasons: i) the measurement set of process variables used did not contain enough information on these types of abnormal events, and ii) the normal operating region was not defined well enough due to a) inadequate sampling of representative operations in the reference set or b) the presence of some mis-classified samples in the reference set, or c) that there really appears to be no stable steady-state reference set.

The SPE plot used in conjunction with the T planes not only helped to signal abnormal operations but it also gave an indication of continuous model mismatch and the need to update the reference data set and model.

7.1.3 Creating Meaningful SPC Planes

As noted earlier, different combinations of T score plots might yield more meaningful monitoring planes. For instance, figures 7.5 and 7.6 show that vectors T3 and T5 yield a valuable plane in which the samples 432-502 (representing low catalyst activity) can be easily spotted.

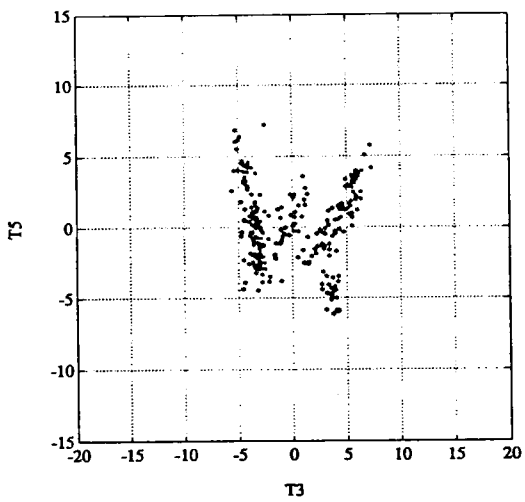


Figure 7.5: Reference SPC T3-T5 monitoring plane.

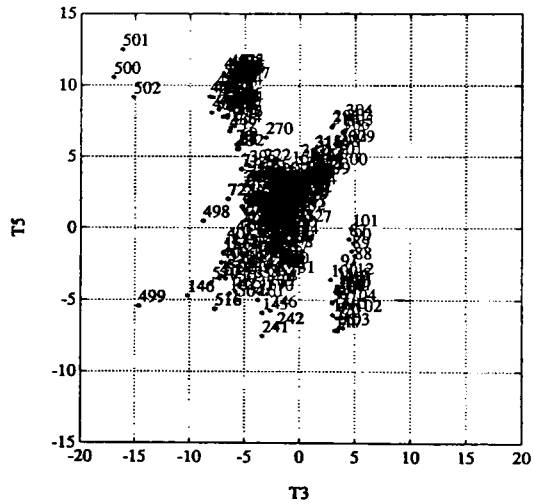


Figure 7.6: Test data plotted in SPC T3-T5 monitoring plane.

In figure 7.6, the sample group of low catalyst activity lie distinctly in the upper left-hand region of the plane, separated from the remainder of the test data (shown in figure 7.5). This exemplifies the power of combining different vectors to yield more meaningful visual monitoring spaces.

7.1.4 Searching for the Cause(s) of Abnormalities

As with a Shewart chart, once an abnormally large SPE value or t score has been flagged, a search for an assignable cause (or causes) should be undertaken. Although it is best to conduct this evaluation as soon after detection as possible, a suggested procedure for evaluating this data set is discussed below.

With PLS, the t score of a sample arises from the product of the value of the process variable x and its appropriate w loading. Thus, there are three ways from which a large change in a t score can arise: i) a large change in a process variable whose w loading is relatively small, ii) a small change in a process variable whose w loading is very large, and iii) a somewhat large change in a process variable whose w loading is also relatively large.

Once a large change in a t score is found, the individual products (dx value)*(w loading) could be searched to find those which contribute most strongly to the change in the t score. These x variables would form a subset which would be examined to find a possible cause for the change in the process. For the FCCU data, the following steps were taken.

1) The t scores for each dimension were plotted as a function of time to identify pairs of samples where the change in t was large.

   Changes caused by large breaks in time were not looked at because it was felt that due to the drifting nature of the process operating point, the value of too many variables could change over a large break in the time history.

2) The individual products (dx value)*(w loading) from step 1 were sorted from greatest to least to identify which process variables had moved. Process variables which were either manipulated (MV) or controlled variables (CV) were noted. A large change in a controlled variable should indicate an operational move or deliberate change. A large change in a MV could represent a common disturbance (if the CV for the loop remains relatively constant) or could be indicative of an operational move if the CV changes significantly as well.

3) The process variables' values at the start and end of the interval (and their difference) were examined to see if they substantiated the process change thought to have taken place. The auto-scaled weights of these x variables were also checked to see if their variance had been scaled up during data pretreatment.

Four sets of process moves from the FCCU abnormal data set were examined: 72-73, 158-159, 160-161 and 499-500. Due to the proprietary nature of the process information, it is not detailed here. However, some general observations are discussed below.

All four changes occurred during process transients. Although initially it was felt that the small time lapse of these intervals would help in the identification of key process variables responsible for the moves, it was later realized that a large number of the process variables were probably in a state of change and not yet at new steady-state values. This complicated the interpretation of the dx values. Although some dx values concurred with the process changes known to be taking place, other process variable changes were directionally opposite to expectation, or insignificant or hard to draw any

conclusion from. It would have been better to look at changes from the last stable point of one group to the first stable point of the next group, such as the intervals 70-73 or 497-500.

7.2 Summary of SPC Monitoring Results

The multiple operating points of the FCCU violated the single steady-state assumption of SPC and contributed greatly to the difficulties in defining a normal operating region and in searching for causes of process changes. The analysis showed that the first few latent variables characterized changes and time variations in the process which were common to both the test and reference data. Later LVs were more useful for monitoring the process and were able to flag start-up, process mode changes, high metals content in the feed and catalyst activity changes. The SPE values, indicating model mismatch suggested how quickly the FCCU reference data might have to be updated to reflect the constantly changing operating point.

If the process had a singular operating point, one might have been able to extend the interpretative value of the manipulated (MVs) and controlled variables (CVs) by building reference models using only the MVs in the X space and CVs in the Y space. Checking for large changes in the CVs would allow one to keep an eye on the model fit and monitoring the MVs would indicate the presence of disturbances (which may or may not be compensated for by the control strategy in place, depending upon whether or not the Y space is disrupted as well). The MV and CV data could also have been split into two separate X blocks and analyzed using hierarchical PLS, as discussed in chapter 8.

CHAPTER 8: HIERARCHICAL PLS ANALYSIS

This chapter contains the results from a hierarchical PLS analysis of the large (1170 sample) data set. It also discusses some differences between the algorithm of Wold et al. (1987b) and the one proposed by Wangen and Kowalski (1988) and how the latter might apply to the FCCU.

8.1 Hierarchical PLS (HPLS)

For the hierarchical (HPLS) case, the Wold et al. approach (1987b) was applied to the data used in the PLS analysis. The X space was broken up into six blocks according to the structure of the unit: i) feed, ii) reactor, regenerator and air blower (called 'RR'), iii) desuperheater and main fractionator ('Frac'), iv) rectified absorber ('RA'), v) debutanizer, and vi) depropanizer. In order to compare the effect of breaking up the X data on the model, a single Y space was maintained for this analysis. Auto-scaling was applied to all variables, consistent with the PLS analysis.

Table 8.1 lists the percentage sum of squares (%SS) explained for each block per model dimension; cross-validation has not yet been implemented for hierarchical PLS, so CSV/SD values are not available. Table 8.2 contains the consensus loadings (v and w vectors) which represent the relative contributions of individual t_a and u_b vectors to the consensus vectors t and u respectively. A relatively large loading means that a block's t_a vector dominates the consensus vector, causing the model to explain a

relatively large portion of that block's variation. The consensus loadings provide basically the same information as the %SS explained values. Both are difficult to interpret directly as each block contains a different amount of the initial SS in X, however, the Wold et al. algorithm adjusts for this by dividing each block by the number of variables contained in it. This gives each t_a vector the same importance in the consensus matrix T although a X_a block may contain only one process variable and others may contain several. The same holds true for the u_b vectors collected in the consensus matrix U.

Table 8.1--HPLS %SS Explained Per Block for Each Latent Variable (X=1170 x 136 total) (Y=1170 x11)

	Feed	RR	Frac	RA	Debut	Deprop	Y	Y Cumul.
Initial SS per block	19873	54943	52605	12859	8183	8183	12859	
X Variables per block	17	48	46	11	7	7	11	
LV: 1	33.5	21.3	25.0	40.1	24.6	26.7	25.8	25.8
2	2.8	10.8	7.0	6.4	43.7	16.9	12.6	38.4
3	9.9	9.5	17.1	12.4	17.6	7.4	12.2	50.6
4	4.2	12.7	7.7	6.8	1.8	3.4	6.9	57.5
5	0.8	0.5	0.5	1.0	0.4	15.8	0.3	57.8
6	1.8	2.3	3.1	7.5	3.2	8.1	10.0	67.8
7	1.9	1.0	2.1	1.5	1.8	8.4	2.2	70.0
8	1.4	2.0	11.5	2.8	0.6	0.4	3.8	73.8
9	4.6	1.6	6.4	1.1	0.4	0.7	1.1	74.9
10	0.6	1.0	1.8	11.8	0.3	0.2	1.0	75.9
11	6.1	3.7	0.4	0.6	0.2	0.4	0.9	76.8
12	5.7	1.0	0.4	0.6	0.0	0.1	0.4	77.2
13	0.4	4.5	0.7	0.1	0.1	0.1	0.3	77.5
14	0.5	0.6	0.5	0.7	0.5	6.3	0.4	77.9
15	5.4	0.7	0.4	0.2	0.0	0.0	0.3	78.2

Table 8.2--HPLS Consensus Loadings Per Block (v and w values)

LV	Feed	RR	Frac	RA	Debut	Deprop	Y
1	246.4	136.5	216.3	265.7	232.3	213.2	302.1
2	21.4	101.6	54.4	60.8	356.0	162.3	147.4
3	75.8	71.2	143.0	122.5	148.4	61.7	142.1
4	28.5	97.2	60.7	54.7	13.6	22.4	80.5
5	-0.1	0.0	-0.2	0.6	0.0	4.5	3.2
6	16.5	14.2	18.1	61.3	36.0	60.5	116.5
7	7.6	6.7	13.1	1.8	10.5	23.2	26.2
8	9.5	15.5	69.8	12.9	3.5	2.4	44.4
9	8.3	6.8	18.5	3.9	1.1	2.1	13.3
10	2.8	5.8	6.4	33.0	1.8	-0.1	12.0
11	18.3	16.0	1.6	2.3	1.2	1.0	11.0
12	4.5	0.7	0.1	0.4	0.0	0.0	5.0
13	0.4	5.6	1.6	0.4	0.4	0.3	4.0
14	2.4	4.0	2.2	2.6	1.1	17.9	4.8
15	2.0	1.0	0.4	0.1	0.1	0.0	3.4

To overcome this interpretation problem, the values of the %SS explained for each block in table 8.1 were multiplied by the fraction of total initial SS that each block contributed to X. These new values were called the "%SS explained of the total X variation", and are presented in table 8.3. This allows one to clearly see from which block the SS in X are being explained per latent variable; it also allows summation of the total %SS of X explained per dimension. The hierarchical results were then compared to the PLS statistics, with which they are listed in table 8.4.

Table 8.3--HPLS %SS Explained of Total X Per Latent Variable

% of Initial Total SS	Feed	RR	Frac	RA	Debut	Depro	Total %SS X	Cumul. Total %SS X
	12.7	35.1	33.6	8.2	5.2	5.2		
LV: 1	4.3	7.5	8.4	3.3	1.3	1.4	26.1	26.1
2	0.4	3.8	2.4	0.5	2.3	0.9	10.2	36.3
3	1.3	3.3	5.7	1.0	0.9	0.4	12.7	48.9
4	0.5	4.5	2.6	0.6	0.1	0.2	8.4	57.3
5	0.1	0.2	0.2	0.1	0.0	0.8	1.4	58.7
6	0.2	0.8	1.0	0.6	0.2	0.4	3.3	62.0
7	0.2	0.4	0.7	0.1	0.1	0.4	2.0	63.9
8	0.2	0.7	3.9	0.2	0.0	0.0	5.0	69.0
9	0.6	0.6	2.1	0.1	0.0	0.0	3.4	72.4
10	0.1	0.4	0.6	1.0	0.0	0.0	2.0	74.4
11	0.8	1.3	0.1	0.0	0.0	0.0	2.3	76.7
12	0.7	0.4	0.1	0.0	0.0	0.0	1.3	78.0
13	0.1	1.6	0.2	0.0	0.0	0.0	1.9	79.9
14	0.1	0.2	0.2	0.1	0.0	0.3	0.9	80.7
15	0.7	0.2	0.1	0.0	0.0	0.0	1.1	81.8

Table 8.4: Comparison of PLS and HPLS Statistical Results (Not cross-validated)

	PLS:				HPLS:			
LV	% SS X	Cumul. % SS X	% SS Y	Cumul. % SS Y	%SS X	Cumul. %SS X	% SS Y	Cumul. % SS Y
1	26.5	26.5	30.4	30.4	26.1	26.1	25.8	25.8
2	13.0	39.5	14.3	44.7	10.2	36.3	12.6	38.4
3	13.4	52.9	8.7	53.4	12.7	48.9	12.2	50.6
4	5.6	58.5	8.7	62.1	8.4	57.3	6.9	57.5
5	4.2	62.7	6.7	68.8	1.4	58.7	0.3	57.8
6	4.0	66.7	4.5	73.3	3.3	62.0	10.0	67.8
7	3.9	70.6	2.5	75.8	2.0	63.9	2.2	70.0
8	1.6	72.2	3.5	79.3	5.0	69.0	3.8	73.8
9	2.8	75.0	1.2	80.5	3.4	72.4	1.1	74.9
10	1.8	76.8	1.1	81.6	2.0	74.4	1.0	75.9
11	1.5	78.3	1.1	82.7	2.3	76.7	0.9	76.8
12	1.1	79.4	1.2	83.9	1.3	78.0	0.4	77.2
13	1.5	80.9	0.7	84.6	1.9	79.9	0.3	77.5
14	1.0	81.9	0.9	85.5	0.9	80.7	0.4	77.9
15	0.9	82.8	0.7	86.2	1.1	81.8	0.3	78.2

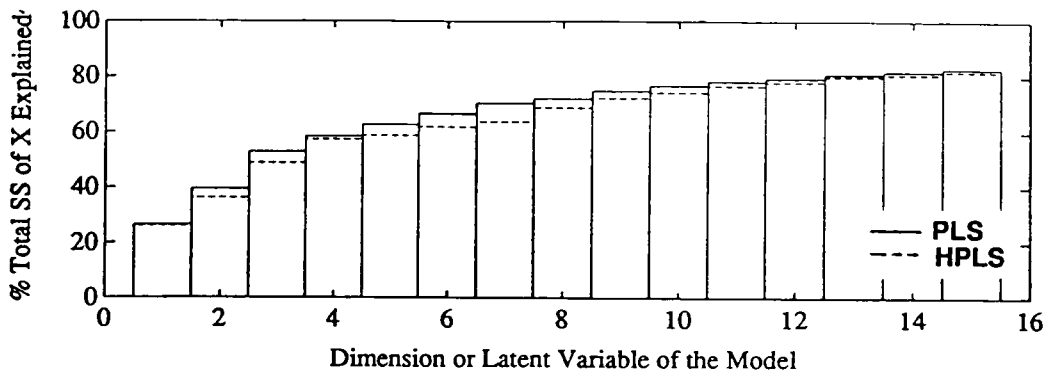


Figure 8.1: Total percentage SS of X explained by PLS and HPLS.

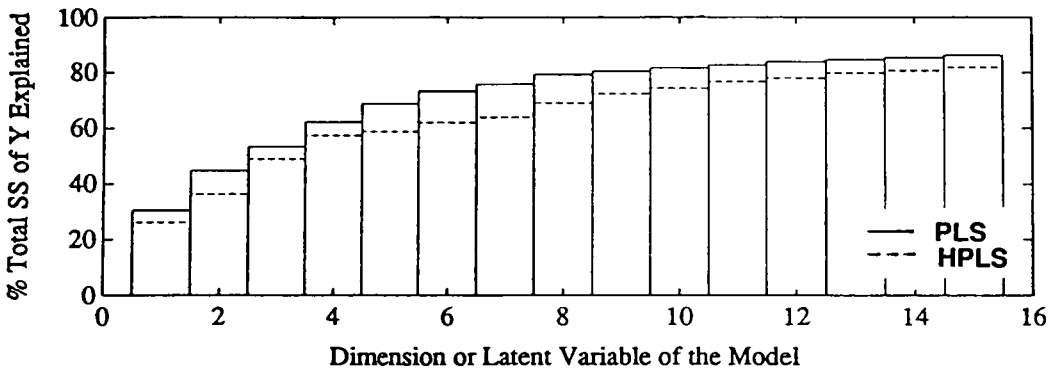


Figure 8.2: Total percentage SS of Y explained by PLS and HPLS.

Figures 8.1 and 8.2 show the percentage of total SS explained of X and Y for both the PLS and HPLS cases. The HPLS case consistently explained slightly less of the Y space while modelling almost the same amount of the X data as the PLS case. Plots of

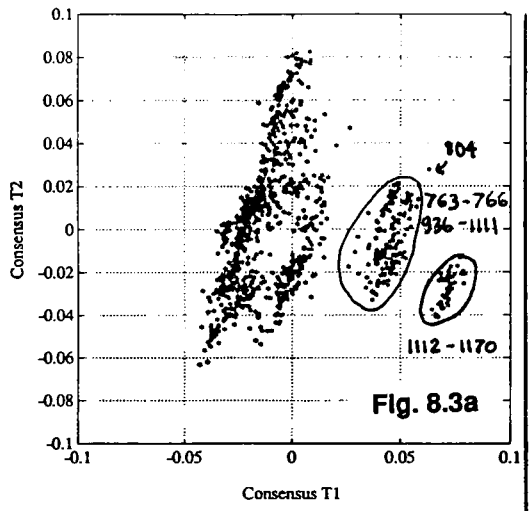
the consensus vectors t and q vectors are shown in figures 8.3a-g and 8.4a-g, repectively. They are also described below in table 8.5. Since no CSV/SD values were available, the Q plots provided the only other information to aid in determining what each LV or dimension was modelling. As was found in the previous analyses, caution must be used here since a noisy product can dominate a q vector although its variation may not be well predicted (normally indicated by a high CSV/SD value).

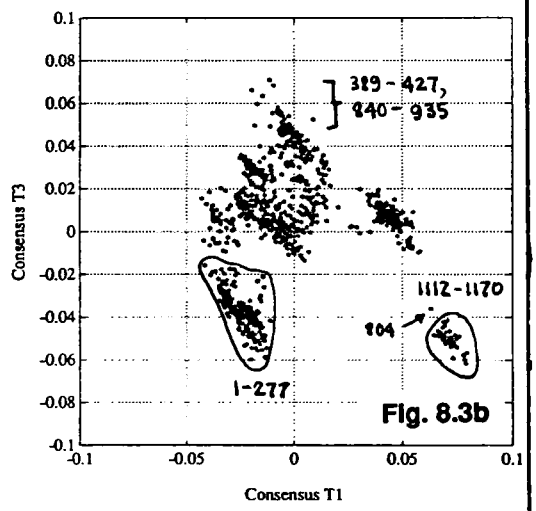
Table 8.5--Analysis of X Space Consensus T Vectors and Y Space Q Vectors

X Space	Y Space
T1: Breaks the process samples up into three distinct groups: 763-766 and 936-1111, 1112-1170, and the rest. It also appears confounded with the time trend.	**Q1:** The plot of the Y products in the Q1-Q2 plane is almost identical to that from the PLS analysis. Q1 separates the desirable gasoline product (4 and 10) from the undesirables fuel gas (1) and coke (8 and 11), as well as LGO (5).
T2: No distinct groupings are modelled, thus the vector is probably describing process variation common to most samples.	**Q2:** The lightest and heaviest low quality products fuel gas (1) and C.O. (7) dominate one end, while butane yield (3) dominates the other. It is similar to the Q2 vector from the PLS analysis.
T3: 1-277, 1112-1170: Are separated from the rest of the data. The layout of the samples in the T1-T3 plane is very similar to the layout in the PLS T1-T3 plane.	**Q3:** Q3 is pretty much a negative version of Q3 from the PLS analysis. Gasoline (4), LGO (5) and IGO (6) dominate the vector, suggesting that the LV is modelling changes at the fractionation unit.

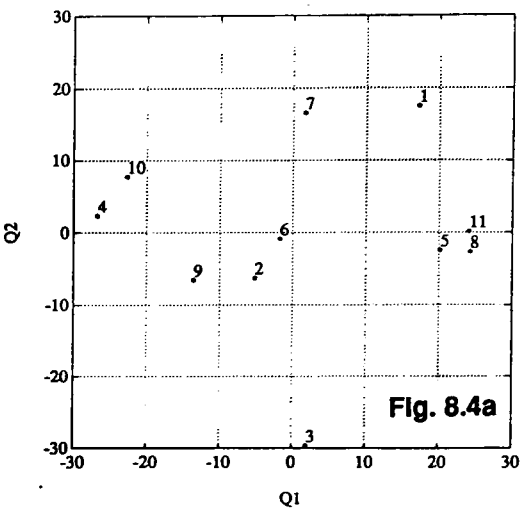
Table 8.5--Continued

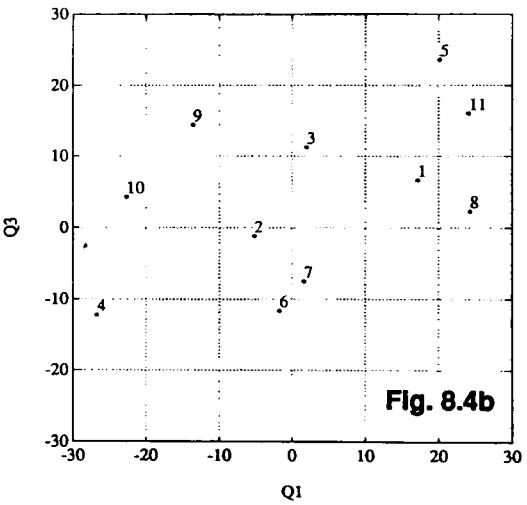
X Space	Y Space
T4: Further separation of the samples takes place: 1-7, 767-823: Low metal feed and boomer (asphalt) operation. 1112-1170: Most recent samples.	**Q4:** Although IGO (6) dominates one end of Q4, it is known to be a noisy variable and may not be highly predictable.
T5: The scattering of samples along the T5 vector reflects its low modelling contribution.	**Q5:** The marginal modelling power of LV 5 is reflected in the Q5 vector where all products lie within a tight range about the zero axis point.
T6: This vector draws apart samples 936-1066 and 1067-1111.	**Q6:** Although the distinction is not clear, Q6 appears to separate the desirable products propane (2) and gasoline (4)4 (along with the desirable yields 9 and 10) from the undesirable products fuel gas (1), C.O. (7) and coke (8 and 11).
T7: The sample group 416-429 lies at one end of the vector while a mixed collection of other samples stretch in the opposite direction.	**Q7:** The low predictive power of this dimension is reflected in the flat plot of the Y products in this plane.
T8: No new separation of points is revealed by this consensus vector.	**Q8:** IGO (6) dominates the Q8 vector; the dimension explains predominantly the SS lying in the fractionation block.



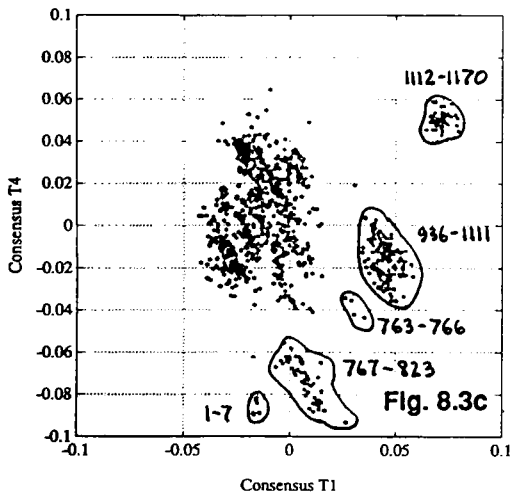


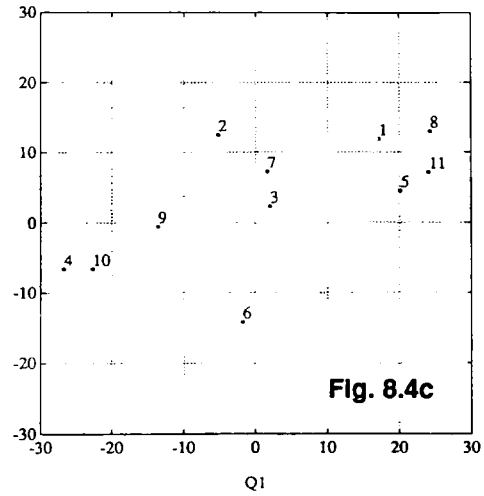
Figures 8.3a-b: Consensus T planes from HPLS.

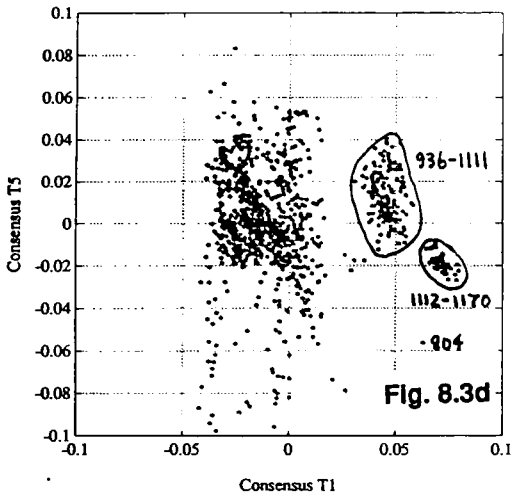


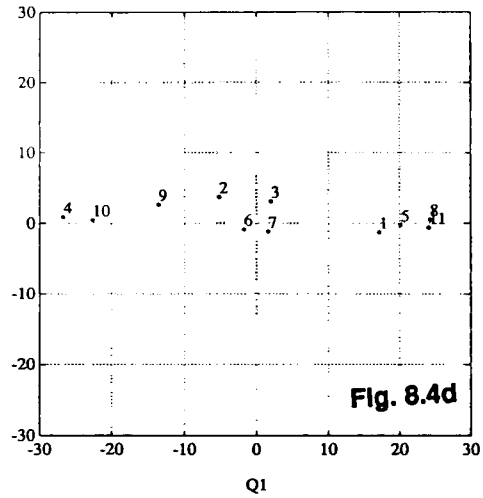


Figures 8.4a-b: Q (product variable) planes from HPLS.



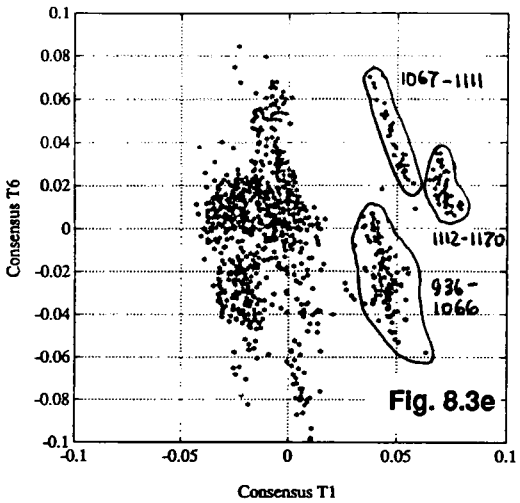


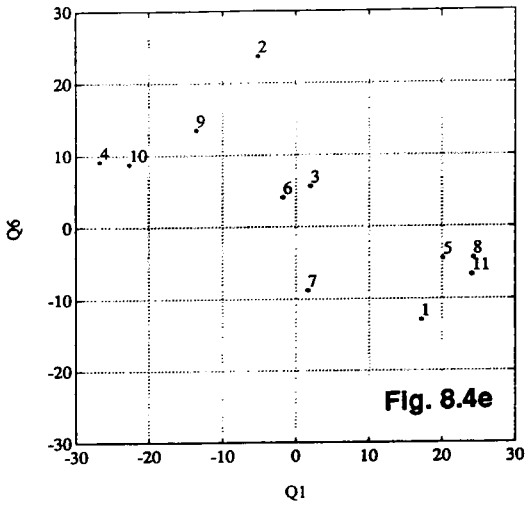


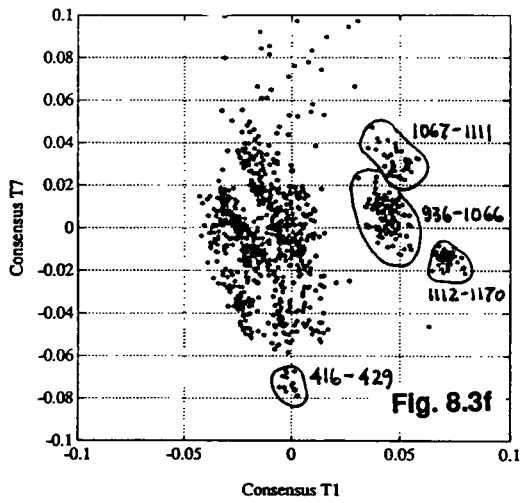


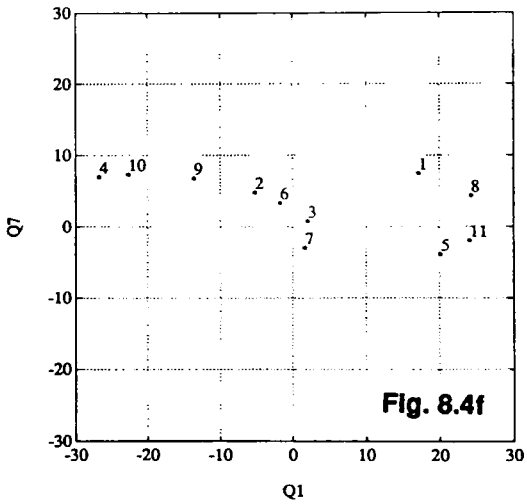
Figures 8.3c-d: Consensus T planes from HPLS.

Figures 8.4c-d: Q (product variable) planes from HPLS.









Figures 8.3e-f: Consensus T planes from HPLS.

Figures 8.4e-f: Q (product variable) planes from HPLS.

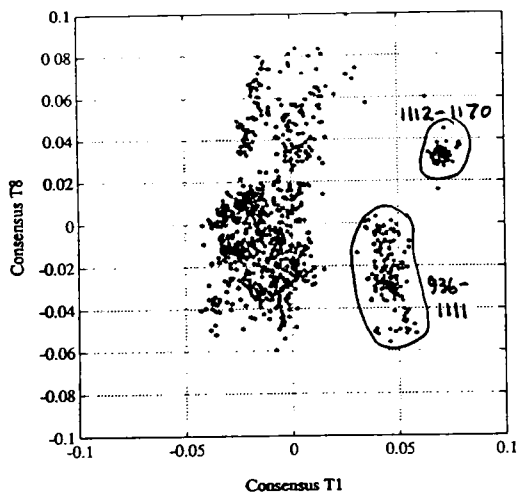
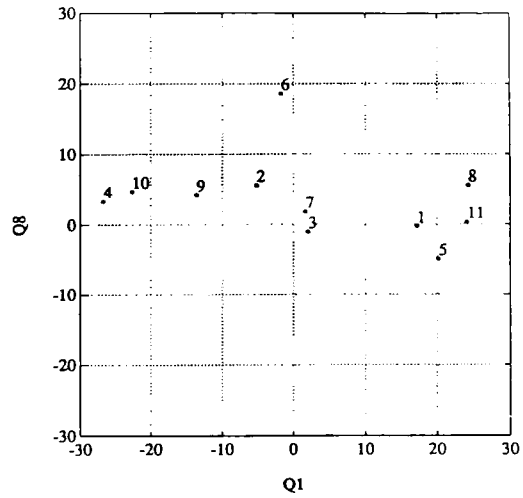


Figure 8.3g: Consensus T planes from HPLS.



Figures 8.4g: Q (product variable) planes from HPLS.

8.2 Discussion of HPLS Results

The fact that the HPLS approach was slightly less powerful than ordinary PLS in predicting the X and Y spaces for this data and that no one block tended to dominate the first few LVs was probably a reflection of the highly coupled nature of the process. If the t_a vectors of a particular dimension described process changes that affected all parts of the unit, the consensus vector would give almost equal weight to each block. This was the

case for the first dimension where five of the six blocks had a consensus loading between 216 and 265. The plots revealed that this dimension was heavily confounded with the time trend and a slowly decreasing unit feed rate.

The second dimension showed more contrast with consensus loadings ranging from 21 to 356. The debutanizer block dominated this dimension, followed by the depropanizer and reactor-regenerator blocks. As the process changes that affected many parts of the process were modelled and removed from the data blocks, the later dimensions focused on subtler changes and modelled SS from predominantly one or two blocks, as happened with the fourth dimension (which described a large portion of the SS in the reactor-regenerator and fractionation blocks). The eighth and ninth dimensions explained mostly SS of X from the fractionation block.

The Wold et al. algorithm assumed that the individual blocks were independent of each other but this was not true for the process data. Although the data was split up in a logical manner according to the process structure, there was still a high degree of correlation amongst the data, most prominently in the direction of the unit flows. Upstream units would be expected to have a significant effect on the downstream portions of the unit.

8.3 Alternate HPLS Approach

Thus, the data from the FCCU was not well suited for this particular algorithm. However, Wangen and Kowalski (1988) modified the Wold et al. algorithm to accommodate more complicated arrangements of blocks, such as cascading structures where some blocks are both predictors and predicted. Such flexibility would allow a more appropriate model structure for the FCCU data, such as the one shown in figure 8.5.

A key issue with this algorithm involves how the correlational information from successive X blocks is carried through to predict the Y space. It appears that block X_{a-1} must be strongly correlated to the next block X_a, in order that information about the correlation between X_{a-1} and the Y space be passed through. Otherwise, this information is lost. Also, since Wangen and Kowalski do not give proof of the vector properties, orthogonality of the intermediate vectors is not clear. The reader is referred to the paper for a detailed discussion of the algorithm.

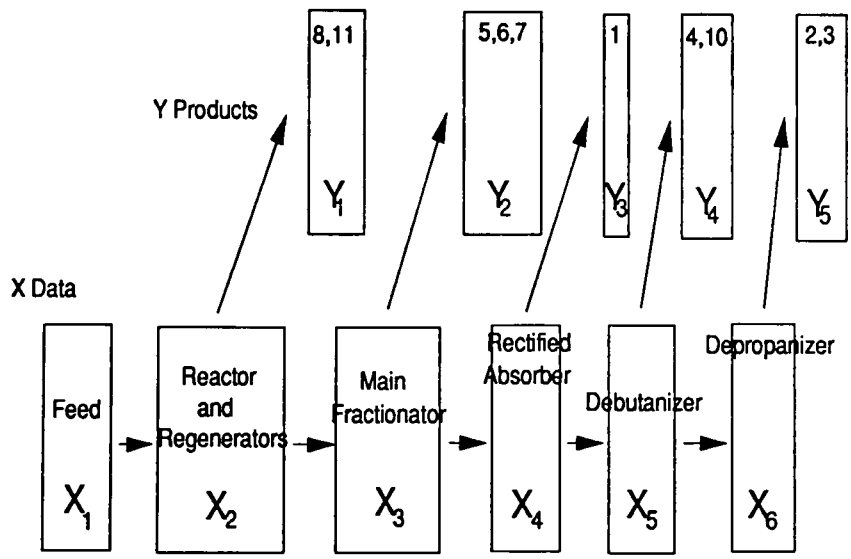


Figure 8.5: Alternative structure for HPLS analysis (using Wangen and Kowalski algorithm) of the FCC Unit.

To illustrate the difference between the two approaches, calculations for one latent variable from a data set consisting of two X blocks and one Y block are shown in figures 8.6 and 8.7.

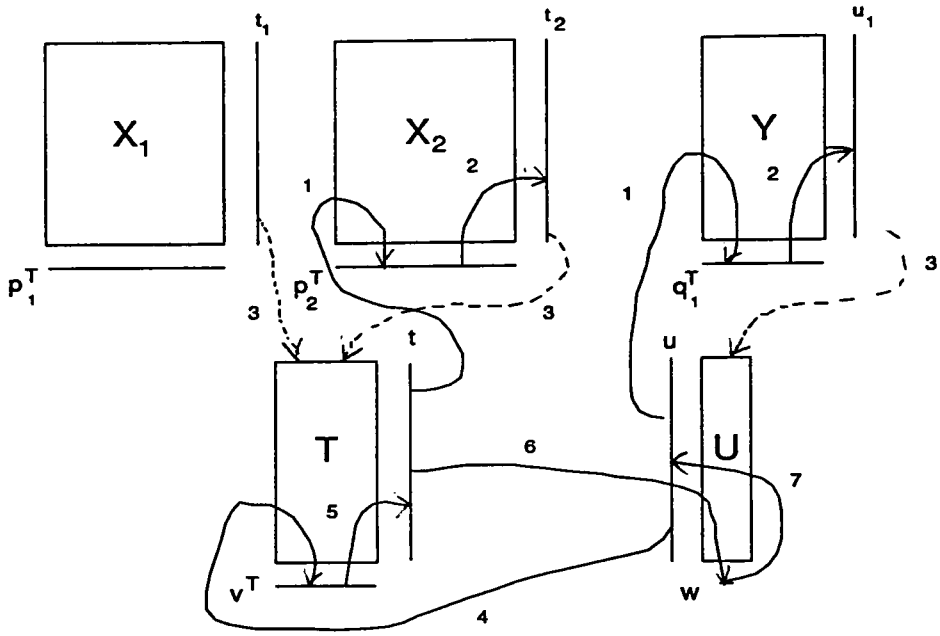


Figure 8.6: Wold et al. HPLS algorithm for a single dimension. The numbers represent the order of the calculations.

The Wold et al. algorithm (figure 8.6) uses the consensus vector t to calculate the individual block t_a score vectors and then performs a NIPALS-PLS round between the X space consensus block T and the Y space consensus block U (which, for the single Y block, is also the u_b vector of the Y block). Predictions for each block are calculated as follows:

$$E_a = X_a - tp_a^T \tag{8.1}$$

$$F = Y - \hat{u}q^T \tag{8.2}$$

$$= Y - btq^T \tag{8.3}$$

(when q is not normalized, b equals unity)

Note that $\hat{u}$ represents the estimate of the u vector and is equivalent to bt.

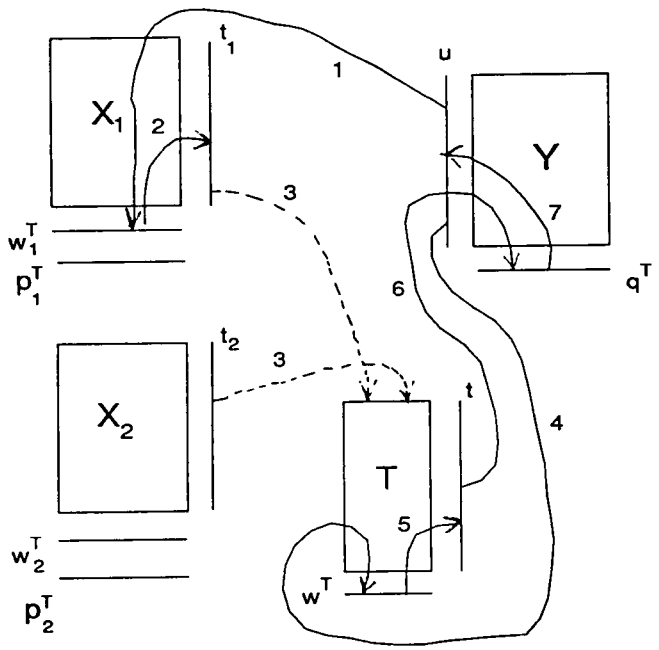


Figure 8.7: Wangen and Kowalski HPLS algorithm for a single dimension. The numbers represent the order of the calculations.

The Wangen and Kowalski approach (figure 8.7) performs regular PLS between the Y space and each X block to get the individual t_a score vectors. These are then collected into the T matrix where the u_b vector from the Y space is used in a PLS round to calculate a consensus score vector t, followed by an update of u. The predictions for each block are calculated as follows:

$$E_a = X_a - tp_a^T \tag{8.4}$$

$$F = Y - \hat{u}q^T \tag{8.5}$$

$$F = Y - (b_1 t_1 + b_2 t_2)q^T \tag{8.6}$$

Here, the coefficients b_a act as weights which determine the contribution of their respective t_a vectors in the model of the Y space.

The t_a vectors of the Wangen and Kowalski approach are calculated so that they correlate strongly to the Y space whereas the Wold et al. t_a vectors are calculated to describe their individual blocks most prominently. Also, since the Wangen and Kowalski t_a vectors remain orthogonal (when there are no intermediate blocks in the X space), they would be suitable for building monitoring planes for all or parts of the process unit.

CHAPTER 9: CONCLUSIONS

The goals of this work were to attempt to use the techniques of PCA and PLS to study the process history of a fluidized catalytic cracking process. Specifically, it was hoped the analyses would reveal interesting periods in the process history, identify interesting relationships amongst the process variables being collected, provide a predictive model of the product space based on operating variables, and provide a fault detection or monitoring space from which process changes and disturbances could be detected.

Both the PCA and PLS analyses revealed periods of process operation characterized by high and low feed rates, high and low feed metal contents, and swings between boomer (asphalt) and non-boomer (regular) operations. These events correlated strongly to the first few latent variables (LVs) extracted. Later LVs highlighted transients which would be obvious to operations personnel (e.g., unit start-ups, intermittent drops in feed rate, preparation for operational changes). The analyses confirmed that the plant was slowly shifting during the time period studied and that process conditions at the start of the time history were distinctly different from those at the end of the time history.

PC analysis of the product space, however, showed that despite these regular process shifts, the plant was producing its products within the same output "window" or plane.

Interpretation of the T, Q and T-U plots from PLS was complicated by the undesigned nature of the process history. Many significant process changes occurred at once (e.g., feed quality, feed rate and temperature changes during moves into or out of boomer operations). Thus, individual phenomena could not be observed. Although the first or second LV correlated strongly with key process changes or events, later LVs could not be expected to do so since they were forced to be perpendicular to each other whereas the process events were not.

The search for causes of detected changes and evaluation of individual process variables' contributions to the models was difficult due to several factors: the large number of process changes that occurred at once, the highly coupled nature of the process, the number of acceptable operating points for the FCCU, non-linearities of the process which made accurate modelling difficult, and the undesigned nature of the data.

Product yield and selectivity models were developed using PLS on as wide a range of operating conditions as possible. The confidence intervals for the eleven LV model (calculated from the fitted prediction errors) were 9-14% of the individual y variables' ranges, and predicted 81.3% (cross-validated) of the sum of squares (SS) in the Y space. This indicated that the numerous operating points of the FCCU were easily captured by the linear PLS model.

In developing an SPC monitoring space, it was found that the first four latent variables of the reference model described process variation common to both the reference and test samples. These dimensions characterized deliberate changes and time variations in the process and so were not practical for monitoring purposes. However, the later latent variables were more sensitive to the differences presented by the test samples. Abnormal events flagged by the T planes or SPE values (or both)

included start-up and process preparation for boomer and non-boomer operations (which caused feed rate and feed quality changes), high metals content in the feed, and catalyst activity changes.

Success of the SPC monitoring space was hampered by the fact that the FCC process did not operate near one set of steady-state conditions, but rather was continuously changing over time. This application of PLS (or PCA) would be much more amenable to processes with a stable operating point (such as would be found in a quality control situation).

For the HPLS analysis, little difference was found between its predictive abilities and interpretation and those of PLS. Statistically, the HPLS analysis modeled slightly less of the X and Y spaces than PLS and the plots revealed fewer subtle changes in the data. The opportunity to gain further insight by seeing which portions of the X data contributed to each dimension in HPLS was limited by the high correlation amongst the blocks.

Some minor conclusions and recommendations for further work are outlined below.

The appropriateness of auto-scaling the data sets was also considered. Variables which, due to their very small variation were scaled up in importance by auto-scaling, did not unduly dominate the models and thus it was felt that the auto-scaling approach was suitable for a first look at this industrial data set.

Unresolved issues specific to the FCCU were: the inability to directly monitor the key process variables of feed quality and catalyst quality on a timely basis, the presence of both fast responding and slow responding process variables in the unit, the high correlation amongst key process changes which made isolation of causes difficult, and having y product variables exiting the unit in the middle of the X data set (i.e., complications imposed by the physical structure of the process).

A performance evaluation of the SPC monitoring space (e.g., number of false alarms and delays in detection) was not undertaken due to the timely nature required for verification. This would require a good model and implementation of the technique on site. Validation of sample classification would also be important for interpretation and for selecting reference data for monitoring spaces.

Robustness to sensor failure should also be considered. Obviously, the model prediction would be sensitive to failure of sensors which are heavily weighted in a PCA or PLS model. Kresta, Marlin and MacGregor (1991c) showed how PLS could easily handle sensor failures.

Overall, it was very apparent that the quality of the data used and the purpose of the analyses were paramount to the successful application of PCA and PLS. Post analysis of the FCCU data revealed interesting changes in the process history but diagnosis of the causes for these changes requires information not provided by the techniques. PLS was quite successful in modelling product yields and selectivities spanning a wide range of operating conditions. The development of multivariate SPC monitoring procedures was less successful, due mostly to the FCCU's continually changing process state. This latter use of PCA and PLS would be much more amenable to processes with a stable operating point (such as would be found in a quality control situation).

APPENDIX A: CHECKING FOR THE NEED TO TIME SHIFT DATA

A.1 Purpose

The main problem with the presence of time delay in a PLS or PCA data set was discussed in section 4.6.2. To re-iterate, when a process is at steady-state, the relationship amongst the process variables and the product qualities should remain relatively constant. However, when disturbances enter the process or operational moves are made, the steady-state relationships are disrupted adding additional but dynamic information to the data set. Such transition periods can have a serious effect on a steady-state model if the time lag is large and the amount of dynamic data relative to steady-state data is great.

If the Y space is shown to be dependent upon past as well as present values of X, then the X block data may have to be replicated and time shifted back one hour (or as many replicates and shifts as needed to fill in the lag period). It is also possible that only certain blocks of the X space need to be replicated. One does not want to shift more X data than necessary because PCA and PLS will "share the wealth" or distribute the loadings amongst all correlated variables used, yielding smaller loadings per correlated variable as the number of variables in the data set increases. This issue was discussed in section 4.9.

A.2 Auto- and Cross-Correlation Check

The auto-correlation of each variable was first examined to see if there were correlations in time. It is also important to do this before any cross-correlation work since highly auto-correlated series will show much more of a pattern than truly exists. The existence of delayed relationships between the process variables and the outputs (i.e., due to recycle, inventory) was checked by cross-correlating each of the 136 x variables with each of the 11 y variables of interest. All variables were differenced before-hand to remove the effects of non-stationarity from the data (since the correlation functions are only applicable to stationary data).

Since the data set contained fifteen natural breaks in the time series, this translated into approximately 2,200 auto- and 22,400 cross-correlation plots. Very few variables (either X or Y) showed significant auto-correlations for time shifted periods of two lags or more (i.e., two hours or more). Many showed a strong correlation to their previous hour's average, and some y variables showed no significant auto-correlation at all.

The case for cross-correlations was more difficult to assess, due in part to the sheer number of plots to be examined. The analysis involved looking for:

I) groups of x or y variables which consistently showed significant lags for time periods of one hour or greater,

II) consistency in the relationship between x and y variables, throughout the fifteen sub-groups of the time history (e.g., feed variables consistently showing a strong correlation to particular y variables), and

III) consistency amongst the data which show significant lags in relation to the physical structure of the unit (e.g., if the propane yield is correlated to the reactor conditions of two time periods before, the butane product would be expected to be related to the same time period since it is the next heavier product and exits the FCCU from the same distillation column as the propane).

The results of an informal survey of these plots is given in table A.1.

Table A.1: Cross-Correlation Results (by Product)

Product	Cross-Correlation Observations
Fuel (dry) gas (1)	This lightest component did not show any significant non-zero lags. Its residence time in the FCCU is very short and thus past values of the process variables are not needed to model this product.
Propane (2) and Butane (3)	Both showed some strong lags to past data from the reactor and both regenerators. The consistency in the lags shown by these two products was expected since they exit the FCCU at the same column (depropanizer) and are similar in nature (both light gases).
Light (5) and Heavy (7) Cycle Oil	Both showed positive lags with data from the reactor and the two regenerators. Product 7, the heaviest of the products, showed some strong correlation to data from two hours previous.
Intermediate Cycle Oil (6)	This product showed no strong correlation with any x variables. This was expected since this product stream is allowed to fluctuate quite widely compared to the others.
Coke Yield (8) and Selectivity (11)	These showed no consistent lags; perhaps the time periods being studied were not long enough to reveal any correlation.
Liquid (9) and Gasoline (4) Yield, and Gasoline Selectivity (10)	The two yields correlated strongly to a scattering of data from the fractionator and downstream columns, but definitely not with the upstream reactor or regenerator as did the light gas products.

A.3 Testing a Time-Shifted Data Set

Based on the above results, it was decided that X data from the feed, reactor, regenerators, air blower and desuperheater would be replicated and shifted back one time period (one hour). This alone increased the size of the X block by 62.5%. If this augmented data set is beneficial for modelling purposes, it is expected that a greater percentage of the Y space will be explained per component (or within the first few components) than with the normal data set.

To test this, two runs were conducted; the first using the normal PLS data set of X (1469 samples by 136 variables) and Y (1469 samples and 11 yields), the second using the augmented X space (1450 samples by 136 variables plus 85 time-shifted variables) and the corresponding Y space (1450 by 11).

A.4 Results

Tables A.2 and A.3 contain the statistical results from the PLS analyses of the normal and augmented data sets, respectively.

The augmented data set actually explained slightly less of the Y space than the normal data set. Comparison of sample space plots (T versus T) and product space plots (Q versus Q) revealed little difference between the two models for the first six components. The closeness of the above results suggested that the time shifted data (t-1 points) were highly correlated to the present t points and that few radical dynamic events

were occurring (where the predictive power of the t-1 data would be evident in boosting the percentage sum of squares explained in the Y space). Expansion of the X space lead to no real improvement in the modelling of Y.

Table A.2: Normal PLS Run Statistics (X=1469 x 136) (Y=1469 x 11)

	Ordinary		Cross-validated Cumul.		
LV	% SS X	% SS Y	% SS Y	% SS Y	Overall CSV/SD
1	22.0	35.8	35.8	35.8	0.801
2	15.8	12.6	12.5	48.3	0.897
3	13.7	7.3	7.2	55.5	0.928
4	4.7	8.6	8.4	63.9	0.900
5	3.6	6.0	5.6	69.5	0.918
6	3.2	4.8	4.6	74.1	0.920
7	4.9	1.7	1.5	75.6	0.970
8	3.3	1.6	1.3	76.9	0.972
9	2.0	2.1	1.5	78.4	0.965
10	1.8	1.7	1.4	79.8	0.962

Table A.3: Augmented PLS Run Statistics (X=1450 x 221) (Y=1450 x 11)

	Ordinary		Cross-validated Cumul.		
LV	% SS X	% SS Y	% SS Y	% SS Y	Overall CSV/SD
1	21.1	35.3	35.3	35.3	0.804
2	17.3	12.0	12.0	47.3	0.903
3	14.3	6.9	6.9	54.2	0.932
4	4.1	8.9	8.8	63.0	0.898
5	4.1	5.1	4.8	67.8	0.933
6	2.8	5.6	5.4	73.2	0.911
7	4.6	1.6	1.4	74.6	0.972
8	3.3	1.7	1.5	76.1	0.969
9	1.6	2.5	1.8	77.9	0.959
10	1.9	1.6	1.1	79.0	0.972

Although this was a crude way to test for the principle, it was felt that the results were clear enough to justify continuing the PCA and PLS work without using shifted data.

REFERENCES

Aastveit A.H., H. Martens, "ANOVA Interactions Interpreted by Partial Least Squares Regression," *Biometrics* **42**,829-844 (Dec.,1986)

Anderson T.W., "*An Introduction to Multivariate Statistical Analysis*", 2nd edition, Wiley, New York (1984).

Arkun Y and G. Stephanopoulos, "Studies in the Synthesis of Control Structures for Chemical Processes: Part IV. Design of Steady-State Optimizing Structures for Chemical Process Units", *AIChE Journal*, **26**(6),975-991(1980).

Bacon D., "*Collection and Interpretation of Industrial Data,*" Queen's University, Kingston, Ontario.

Basseville M., "Detecting Changes in Signals and Systems-A Survey", *Automatica*, **24**(3),309-326(1988).

Box G.E.P., W.G. Hunter and J.S. Hunter, "*Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building*", Wiley, New York (1978).

Box G.E.P., W.G. Hunter, J.F. MacGregor and J. Erjavec, "Some Problems with the Analysis of Multiresponse Data", *Technometrics*,**15**(1),33-51(1973).

Box G.E.P., J.F. MacGregor, "The Analysis of Closed-loop Dynamic-Stochastic Systems," *Technometrics* **16**(3),391-398(1974).

Box G.E.P. and G.M. Jenkins, *Time Series Analysis Forecasting and Control*', Holden-Day, Oakland CA (1976).

Brice J.C. and K.V. Krikorian, "Improve FCC Profitability with Better Control", *Hydrocarbon Processing*, 83-87 (May 1983).

Chatfield C. and A.J. Collins, "*Introduction to Multivariate Analysis*", Chapman and Hall, New York (1980).

Dean R.R., J.L. Mauleon and W.S. Letzsch, "Resid Puts FCC Process in New Perspective", *Oil and Gas Journal*, **80**(40),75-80 (4 Oct 1982).

Dhurjati P., D.E. Lamb and D. Chester, "Experience in the Development of an Expert System for Fault Diagnosis in a Commercial Scale Chemical Process", Proceedings of the First International Conference on Foundations of Computer Aided Process Operations, Park City, Utah, 5-10 July (1987) 589-625.

Draper N.R. and H. Smith, *Applied Regression Analysis*, 2nd edition, Wiley, New York (1981).

Elnashaie S.S.E.H. and S.S. Elshishini, "Industrial Fluid Catalytic Cracking. A Mathematical Modelling Approach", 73rd Cdn. Conference and Exhibition/ 40th Cdn. Chem. Engineering Conference and Exhibition/ 1990 CIC Congress, Halifax 15-20 July 1990.

Errazu A.F., H.I. de Lasa and F. Sarti, "A Fluidized Bed Catalytic Cracking Regenerator Model. Grid Effects", *Cdn. J. of Chem. Eng.*, **57**, 191-197(1979).

Geladi P. and B. Kowalski, "Partial Least Squares Regression: A Tutorial", *Analytica Chim Acta* **185**,1-17(1986).

Geladi P., "Notes on the History and Nature of Partial Least Squares (PLS) Modelling", *J. Chemometrics*,**2**,231-246(1988).

Harris T.J. and W.H. Ross, "Statistical Process Control Procedures for Correlated Observations", *Cdn. J. Chem. Eng.*, **69**, 48-57(1991).

Healy J.D., "A Note on Multivariate CUSUM Procedures", *Technometrics*, **29**,409-412(1987).

Himmelblau D.M., *Fault Detection and Diagnosis in Chemical and Petrochemical Processes*, Elsevier Scientific Publishing Co., New York (1978).

Himmelblau D. M., *Process Analysis by Statistical Methods*, Wiley, New York (1970).

Holly W., R. Cook, and C.M. Crowe, "Reconciliation of Mass Flow Rate Measurements in a Chemical Extraction Plant," *Cdn. J. Chem Eng* **67**,595-601(Aug. 1989).

Hoskuldsson A., "PLS Regression Methods," *J. Chemometrics* **2**,211-228(1988).

Hotelling H., "Multivariate Quality Control" in *Techniques of Statistical Analysis*, eds. Eisenhart C., M.W. Hastay and W.A. Wallis, McGraw Hill, New York (1947) 111-184.

Jackson J.E., "Quality Control Methods for Several Related Variables", *Technometrics*,**1**(4)359-377(1959).

Johnson R.A. and M. Bagshaw, "The Effect of Serial Correlation on the Performance of CUSUM Tests", *Technometrics*,**16**,103-112(1974).

Kemp K.W., "An Example of Errors Incurred by Erroneously Assuming Normality for CUSUM Schemes", *Technometrics*,**9**,457-464(1967).

King R., "Early Detection of Hazardous States in Chemical Reactors", Preprints of IFAC Symposium, Bournemouth, UK 8-10 December (1986) 93-97.

Kraemer D.W., U. Sedran and H.I. de Lasa, "Catalytic Cracking Kinetics in a Novel Riser Simulator", *Chem. Engng. Sci.*, **45**(8),2447-2452(1990).

Kresta J.V., T. Marlin and J.F. MacGregor, "Choosing Inferential Variables Using Projection
to Latent Structures with Application to Multicomponent Distillation", paper 23F, AIChE Annual Meeting Chicago Il., (Nov.,1990).

Kresta J.V., J.F. MacGregor and T.E. Marlin, "Multivariate Statistical Monitoring of Process Performance", *Cdn. J. of Chem Eng* **69**(1),35-47(1991a).

Kresta J.V., T.E. Marlin and J.F. MacGregor, "A General Method for the Development of Inferential Control Schemes Using PLS", Proceedings of the Fourth International Symposium on Process Systems Engineering (PSE), Montebello, Quebec, Aug 5-9 (1991b).

Kresta J.V., T.E. Marlin and J.F. MacGregor, "Development of Inferential Process Models
Using PLS", submitted to *Computers and Chemical Engineering,* (1991c).

Kvalheim O.M., "A Partial Least-Squares Approach to Interpretative Analysis of Multivariate Data", *Chemo & ILS*, **3**, 189-197(1988).

Kvalheim O. M. and T.V. Karstang, "Interpretation of LV Regression Models," *Chemo & ILS* **7**,39-51(1989).

Lee E. and F.R. Groves Jr., "Mathematical Model of the Fluidized Bed Catalytic Cracking Plant", *Trans. Soc. Comput. Simil.* **2**(3), 219-236(1985).

Lee W. and V.W. Weekman Jr., "Advanced Control Practice in the Chemical Process Industry: A View from Industry", *AIChE Journal*, **22**(1), 27-38 (1976).

Lorber A. and B. Kowalski, "A Note on the Use of the Partial Least-Squares Method for Multivariate Calibration", *Appl. Spectroscopy* **42**(8) 1572-1574 (1988).

McDonald G.W.G. and B.L. Harkins, "Maximizing FCC Profits by Process Optimization", presented at the National Petroleum Refiners Association Annual Meeting, San Antonio, Texas. 29-31 March (1987).

McFarlane R.C. and D.W. Bacon, "Empirical Strategies for Open-Loop On-line Optimization", *Cdn. J. Chem. Eng.*, **67**,665-677(1989).

McFarlane R.C., R.C. Reineman, J.F. Bartee and C. Georgakis, "Dynamic Simulator for a Model IV Fluid Catalytic Cracking Unit", Prepared for AIChE Annual Meeting, Chicago, Ill. 14 November (1990).

McGreavy C. and P.C. Smith, "Dynamic Characteristics of the Fluid Catalytic Cracking Process", ISCRE 8, The Eighth International Symposium on Chemical Reaction Engineering, Institute of Chemical Engineers, Edinburgh, Scotland. 10-13 September (1984). Symposium Series No. 87, Pergamon Press.

MacGregor J.F., "On-line Statistical Process Control", *Chem. Eng. Progress*, 21-31(Oct.,1988).

MacGregor J.F., T.E. Marlin, J.V.Kresta and B. Skagerberg, "Multivariate Statistical Methods in Process Analysis and Control", CPC-IV Proceedings of the Fourth International Conference on Chemical Process Control, South Padre Island, Texas, 18-22 February (1991a).

MacGregor J.F., T.E. Marlin and J.V. Kresta, "Some Comments on Neural Networks and Other Empirical Modelling Methods", CPC-IV Proceedings of the Fourth International Conference on Chemical Process Control, South Padre Island, Texas, 18-22 February (1991b).

Mardia K.V., J.T. Kent and J.M. Bibby, *Multivariate Analysis*, Academic Press, Toronto (1979).

Martens H., S. Wold and M. Martens, "A Layman's Guide to Multivariate Data Analysis",in *Food Research and Data Analysis*, eds. Martens H. and H. Russwurm Jr., Applied Science Publ., London (1983).

Martens H., "Multivariate Calibration", PhD Thesis, Technical Univ. Norway, Trondheim (1985).

Martens M., and H. Martens, "NIR Reflectance Determination of Sensory Quality of Peas," *Applied Spectroscopy*,**40**(3),303-310(1986).

Monge J.J. and C. Georgakis, "The Effect of Operating Variables on the Dynamics of Catalytic Cracking Processes", *Chem. Eng. Comm.*, **60**,1-26 (1987).

Moseholm L., "Analysis of Air Pollution Plant Exposure Data: The Soft Independent Modelling of Class Analogy (SIMCA) and Partial Least Squares Modelling with Latent Variables (PLS) Approaches," *Environmental Pollution* **53**(1988) pp.313-331

Nilesh S.K. and P.J. Gemperline, " A Program for Calculating Mahalanobis Distances Using Principal Components Analysis", *TRAC, Trends in Analytical Chemistry*, **8**(10), 357-361(1989).

Palazoglu A. and T. Khambanonda, "Dynamic Operability Analysis of a Fluidized Catalytic Cracker", *AIChE Journal*, **33**(6),1037-1040(1987).

Piovoso M.J., K.A. Kosanovich and J.P. Yuk, "Process Data Chemometrics", paper presented to Chemical Engineering Department, McMaster University, Hamilton, Ontario. 10 May (1991).

Ramesh T.S. and J.F. Davis, "CATCRACKER: An Expert System for Process and Malfunction Diagnosis in Fluid Catalytic Cracking Units", AIChE Annual Meeting, San Francisco, CA, (Nov., 1989).

Shah Y.T., G.P. Huling, J.A. Parakos and J.D. McKinney, "A Kinematic Model for an Adiabatic Transfer Line Catalytic Cracking Reactor", *Ind. Eng. Chem., Process Des. Dev.*, **16**(1),89-94 (1977).

Shum S.K., J.F. Davis, W.F. Punch III and B. Chandrasekaran, "An Expert System Approach to Malfunction Diagnosis in Chemical Plants", *Comput. Chem. Engng*,**12**(1),27-36(1988).

Skagerberg B., "SIMCA", MATLAB 386 Version, McMaster Advanced Control Consortium, Department of Chem. Eng., McMaster University, Hamilton, Ontario (1990).

Upson L.L., "What FCC Catalyst Tests Show", *Hydrocarbon Processing*, **60**,253-258 (Nov., 1981).

Venuto P.B. and E.T. Habib, "Catalyst-Feedstock-Engineering Interactions in Fluid Catalytic Cracking", *Catal. Rev.-Sci. Eng.*, **18**(1)1-150(1978).

Wangen L.E. and B. Kowalski, "A Mutliblock Partial Least Squares Algorithm for Investigating Complex Chemical Systems", *J. Chemometrics*,**3**,3-20(1988).

Willsky A.S., "A Survey of Design Methods for Failure Detection in Dynamic Systems", *Automatica*, **12**, 601-611(1976).

Wise B.M. and N.L. Ricker, "Feedback Strategies in Multiple Sensor Systems," AIChE Symposium Series, Vol. 85, No. 267, 19-23(1989).

Wise B.M., D.J. Veltkamp, B. Davis, N.L. Ricker and B.R. Kowalski, "Principal Components Analysis for Monitoring the West Valley Liquid Fed Ceramic Melter", Waste Management 1988 Proceedings. Tucson AZ Feb 28-Mar 3 1988 811-818.

Wold S., "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models," *Technometrics*, **20**(4), 397-405 (1978).

Wold S., P. Geladi, K. Esbensen and J. Ohman,"Multi-way Principal Components- and PLS-Analysis", *J. Chemometrics* **1**,41-56(1987a).

Wold S., Hellberg S., T. Lundstedt, M. Sjostrom and H. Wold, "PLS Modeling with Latent Variables in Two of More Dimensions", Version 2.1, Frankfurt PLS-Meeting (Sept., 1987b).

Wold S., K. Esbensen and P. Geladi, "Principal Components Analysis," *Chemo & ILS* **2** 37-52 (1987c).

Wold S., M. Sjostrom, R. Carlson, T. Lundstedt, S. Hellberg, B. Skagerberg, C. Wikstrom and J. Ohman, "Multivariate Design," *Analytica Chimica Acta*, **191**,17-32(1986).

Wold S., C. Albano, W.J. Dunn III, K. Esbensen, S. Hellberg, E. Johansson, M. Sjostrom, "Pattern Recognition: Finding and Using Regularities in Multivariate Data", in *Food Research and Data Analysis*, eds. Martens H. and H. Russwurm Jr., Applied Science Publ., London (1983).

Wold S., C. Albano, J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W.Lindberg and M. Sjostrom, "Multivariate Data Analysis in Chemistry", in *Chemometrics - Mathematics and Statistics in Chemistry*, B. Kowalski, Ed., Reidel Publishing Co., Dordrecht, NL(1984).