INDUSTRIAL BATCH DATA ANALYSIS

USING LATENT VARIABLE METHODS

INDUSTRIAL BATCH DATA ANALYSIS USING LATENT VARIABLE METHODS

By

CECILIA PEREIRA RODRIGUES, B.Eng.

A Thesis

Submitted to the School of Graduate Studies

In Partial Fulfillment of the Requirements

For the Degree

Master of Applied Science

McMaster University

MASTER of APPLIED SCIENCE (2006) McMaster University
(Chemical Engineering) Hamilton, Ontario, Canada

TITLE: Industrial Batch Data Analysis Using Latent Variable Methods

AUTHOR: Cecilia Pereira Rodrigues, B.Eng.
(Universidade Estadual de Campinas, UNICAMP)

SUPERVISORS: Dr. John F. MacGregor

Dr. Theodora Kourti

NUMBER OF PAGES: *xi*, 125

Abstract

Currently most batch processes run in an open loop manner with respect to final product quality, regardless of the performance obtained. This fact, allied with the increased industrial importance of batch processes, indicates that there is a pressing need for the development and dissemination of automated batch quality control techniques that suit present industrial needs.

Within this context, the main objective of the current work is to exemplify the use of empirical latent variable methods to reduce product quality variability in batch processes. These methods are also known as multiway principal component analysis (MPCA) and partial least squares (MPLS) and were originally introduced by Nomikos and MacGregor (1992, 1994, 1995a and 1995b). Their use is tied with the concepts of statistical process control (SPC) and lead to incremental process improvements.

Throughout this thesis three different sets of industrial sets of data, originating from different batch process were analyzed.

The first section of this thesis (Chapter 3) demonstrates how MPCA and multi-block, multiway, partial least squares (MB-MPLS) methods can be successfully used to troubleshoot an industrial batch unit in order to identify optimal process conditions with respect to quality. Additionally, approaches to batch data laundering are proposed.

The second section (Chapter 4) elaborates on the use of a MPCA model to build a single, all-encompassing, on-line monitoring scheme for the heating phase of a multi-grade batch annealing process. Additionally, this same data set is used to present a simple alignment technique for batch data when on-line monitoring is intended (Chapter 5). This technique is referred to as pre-alignment and it relies on the use of a PLS model to predict the duration of new batches. Also, various methods for dealing with matrices containing different sized observations are proposed and evaluated.

Finally, the last section (Chapter 6) deals with end-point prediction of a condensation polymerization process.

Acknowledgements

I would like to express my most sincere gratitude towards my supervisors, Dr. MacGregor and Dr. Kourti for all your help and guidance during my studies. I could not have hoped for better professors or friends.

I am thankful for the financial support granted by the Department of Chemical Engineering of McMaster University. Also, I am greatly appreciative of the data and knowledge provided by: MaryAnn Ouimet from Dofasco; Puvin Pichai from Arkema; Flavio Cavalcanti, Maria Vitoria Miron and Paula Ambrogi from Oxiteno.

Special thanks to Salvador Garcia, Kevin and Mark-John for their amity and extensive help in academic matters. Thanks are extended to all those that, through their friendship, made my stay in Canada simply wonderful: Nadira, David D., David L., Art, Jen, Laura, Marta, Chris, Danielle and all the MACC students. I will cherish the memories of our time together always.

I am very grateful to those that were my most admirable professors and guides: Gustavo Valenca, Jose Roberto Nunhez, Sergio Stella and Roberto Gallo.

I would like to thank my father, Nilo, and my sisters, Ligia and Elisa, for their presence in my life. Special thanks to my grandparents, Beatriz and Ortesio da Silva, for all those wonderful years of love and care. Thanks also to all family members and friends who are very dear to me but are too many to list.

Finally, I would like to dedicate this thesis to my mother, Maria Aparecida A. P. da Silva, for setting an example of dedication, strength, fairness and love. Thank you.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The last two decades have witnessed fast changing marketing conditions (Berber, 1995); products are brought in and taken out of the market at a very quick pace. According to Edgar (1996), all areas of manufacturing are seeing more emphasis on the rapid delivery of differentiated products, resulting in smaller plants that are located closer to the customers.

As a direct consequence of this setting, many chemical producers are moving from the relatively stable world of continuous plant production to the more versatile and turbulent environment of multi-product batch production (Bonvin, 1998). This fact, allied with the continuous growth of the net-worldwide product consumption, makes it so that batch and semi-batch processes play an important role in the current industrial environment. There is a strong tendency in the use of these processes to produce high added value substances (specialty chemicals) such as pharmaceuticals, polymers, semiconductors and biochemicals (Nomikos and MacGregor, 1995a).

Today's prevailing situation in the industry with respect to batch processes is characterized by: high final quality and safety requirements, short-time-to-market demands and tight economic investments (Friedrich and Perne, 1995).

For most continuous processes, standard control and optimization techniques can be easily employed to achieve the tight product quality and time demands required by industry. However, technical and operational characteristics inherent to batch processes, such as dynamic and nonlinear behavior, infrequent quality related measurements and

time limited corrective actions, make the aforementioned objectives difficult to achieve (Flores-Cerrillo, 2003 and Bonvin, 1998).

Reduction in product quality variability within a batch process can be accomplished by the successive elimination of sources of the disturbance through various levels of control efforts. These efforts are schematized in Figure 1.1 and described in the following paragraphs.

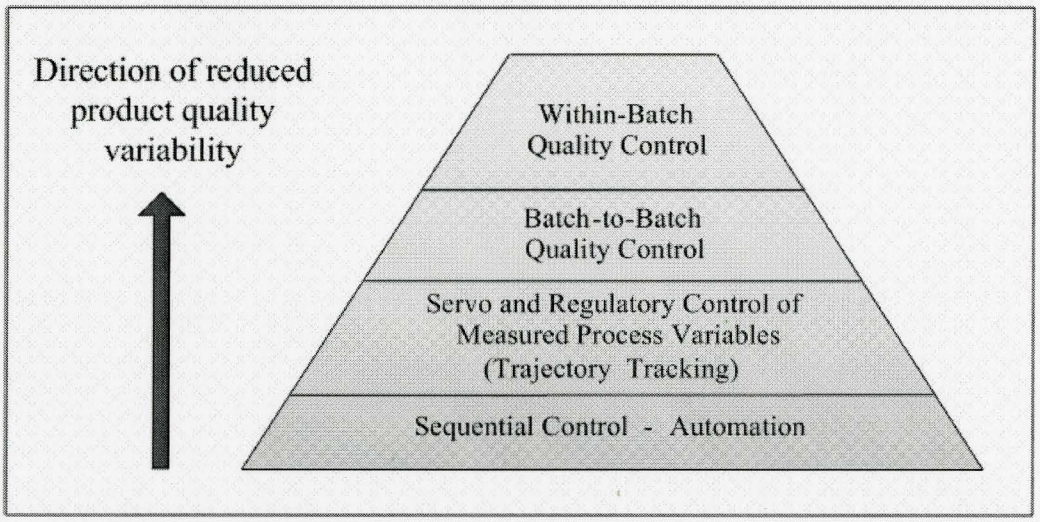

Figure 1.1 Product quality control hierarchy for batch processes.

Once a successful batch production recipe has been developed (normally in a laboratory setting), efforts have to be made in order to ensure that it is followed with as much precision as possible. This can be achieved by automation or sequence control; by coordinating the operations of the equipments required to implement recipe steps, batch-to-batch variations caused by human error can be eliminated (Juba and Hammer, 1986).

Within this logic, there is also need to guarantee that selected process variables are capable of tracking their pre-determined trajectories as tightly as required (lee *et al.*, 1999, Bonvin, 1998, Juba and Hammer, 1986 and Kaparissides and Shah, 1983). Thus, servo-regulatory control of these variables aims at reducing product quality variability by eliminating the effects caused by their deviations from set point values.

Automation of the batch recipe and satisfactory trajectory tracking are very important in reducing variations in product quality (an industrial example is given in Yabuki *et al.*, 2000). None the less, these types of control are not able to directly compensate for quality variability originating from impurities in the raw material, variations in shared utilities or processing problems such as equipment failure, fouling, mixing insufficiencies or lack of available downstream equipment (Bonvin, 1998 and Juba and Hamer, 1986). This can be obtained by what are known as batch quality control techniques.

According to Flores-Cerrillo (2003), batch quality control techniques can be classified depending on their objective (optimization, regulatory control or monitoring) and type of information used (batch-to-batch and within batch).

Batch-to-batch control techniques are capable of reducing variability in the final product quality by using information from previous, completed, batches to make corrections for the next ones (Wiel *et al.*, 1992, Fillipi-Bossy *et al.*,1989, Clarke-Pringle and MacGregor, 1998 and Dong *et al.*, 1996). So, only disturbances that are batch-to-batch correlated can be potentially eliminated through the use of this technique, since it is essentially an off-line control strategy that is only employed between batches. In order to compensate for disturbances in a real time manner, within-batch control techniques must be applied (Kozub and MacGregor, 1992 and Flores-Cerrillo and MacGregor, 2004). Additionally, a set of control techniques that combine information from previous batches to that of the current batch are also available (Lee and Lee, 2003 and Flores-Cerrillo and MacGregor, 2003).

Ultimately, the amount of variability reduction obtained by the implementation of each of these levels of control changes from process to process. Depending on how strict final product specifications are, only the most basic levels of control are required. However, currently most batch processes still only employ automation and trajectory tracking techniques, thus running in an unsupervised manner with respect to quality (Nomikos and MacGregor, 1995a), irrespective of the results obtained. This is mainly

due to the fact that most batch quality control techniques available are time consuming and thus do not suit current industrial needs.

1.1 Thesis Objective

Based on the importance of batch processes in the current industrial setting and on the fact that most of these systems run in an open loop manner with respect to final product quality, irrespectively of the performance obtained, it is concluded that there is a pressing need for the identification and dissemination of quality control techniques that fit present industrial requirements.

Within this context, the main objective of the current work is to exemplify the use of empirical latent variable methods to reduce final product quality variability in batch processes. These methods are also known as multiway principal component analysis (MPCA) and partial least squares (MPLS) and were originally introduced by Nomikos and MacGregor (1992, 1994, 1995a and 1995b).

The use of these latent variable models is tied to the concept of statistical process control (SPC). They are employed to identify, study and eliminate special cause variations, both in an off-line and on-line manner, through the analysis of historical datasets and/or monitoring of current batch trajectories. Corrective actions are determined and executed by plant personnel and can lead to incremental process improvements.

The main advantage of empirical or data-driven models over fundamental ones is that, once the input/output data has been collected, the time required to build such models can be very short. This is especially true in the cases where software packages have been developed precisely for such purpose, as is the case with MPCA and MPLS.

When compared to other empirical methods (i.e. neural networks, ARMAX-type identification, multivariate regressions, among others) the latent variable approach presents the following advantages (Flores-cerrillo, 2003) : i) capability of handling highly correlated data; ii) capability of handling missing data; iii) do not require large training

datasets; iv) are fast, easy to build and update; v) capability of modeling both the Y and X space; vi) provide simple interpretation and data validity tools.

1.2 Thesis Outline

This thesis consists of seven chapters.

Chapter 2 reviews the multiway, multivariate, latent variable methodology proposed by Nomikos and MacGregor (1992, 1994, 1995a and 1995b) .

Chapter 3 demonstrates how MPCA and multi-block, multiway, partial least squares (MB-MPLS) methods can be successfully used to troubleshoot an industrial batch unit in order to identify optimal process conditions with respect to quality. Additionally, approaches to data laundering of time-varying batch process variables are proposed.

Chapter 4 elaborates on the use of a MPCA model to build a single, all-encompassing, on-line monitoring scheme for the heating phase of a multi-grade batch annealing process.

Chapter 5 presents a simple alignment technique for batch data when on-line monitoring is intended. This technique relies on the use of a PLS model to predict the duration of new batches and is again demonstrated on an on-line MPCA monitoring scheme built for an industrial batch annealing process. Additionally, various methods for dealing with matrices containing different sized observations are proposed and evaluated.

Chapter 6 deals with the end-point prediction of a condensation polymerization process.

Chapter 7 contains the main conclusions reached throughout this work.

Chapter 2

Theoretical Description of Multiway Latent Variable Methods

The objective of the current chapter is to review the empirical methodology proposed by Nomikos and MacGregor (1992, 1994, 1995a and 1995b) for the analysis of batch process data. This methodology is based on the use of multivariate latent variable or projection methods, mainly principal component analysis (PCA) and partial least squares (PLS), and supplementary data processing techniques (i.e. trajectory alignment and batch-wise unfolding).

2.1 PCA and PLS

In most industrial plants, massive amounts of process and quality measurements are continuously collected and stored with a frequency that ranges from seconds to hours, respectively. These variables are often highly correlated and prone to having missing values due to the occurrence of process or sensor faults. While ordinary linear regression techniques are incapable of handling data with these characteristics, traditional statistical process control (SPC) methods, such as Shewhart charts, do not take the multivariate nature of the data into account (Kourti and Macgregor, 1995 and 1996 and Undey and Cinar, 2002).

Multivariate statistical methods such as principal component analysis (PCA) and partial least squares (PLS) have proven to be of great use in the analysis of industrial

process data (Eriksson *et al.*, 1999). These techniques are capable of modeling the main variations in the original data set by projecting it onto a lower-dimensional space defined by new variables know as principal components (PCs) or latent variables (LVs). Historical overviews of PCA and PLS and it's applications are given by Wold *et al.* (1987) and Geladi and Kowalski (1986), respectively.

Prior to building any PCA or PLS model it is important that the original data set be mathematically pre-processed in order to transform the data into a form suitable for analysis. Conventional data pre-treatment consists of mean centering and scaling to unit variance. Batch data, however, requires additional processing steps, known as unfolding and alignment, which are described in sections 2.2 and 2.3 respectively.

Mathematically, PCA is represented by (Kourti *et al.*, 1995):

$$X = \sum_{a=1}^{A} t_a p_a^T + E \qquad (2.1)$$

Where: X is the original data set, composed of J variables and I observations; t_a is a score vector and it represents the projection of each observation onto a particular latent variable (a); p_a is a vector of loadings, which expresses the relative importance (weight) of each variable from the original data set to a particular latent variable (a); E describes the matrix of residuals; A is the number of latent variables used in the model.

PLS models incorporate both process (X) and quality (Y) variables and aim at maximizing the covariance between them. These models consist of an outer relation (described by equations 2.1 and 2.2) and an inner relation (described by equation 2.3 for it's most common form, various other models that describe the association between X and Y are presented by Geladi and Kowalski, 1986).

$$Y = \sum_{a=1}^{A} u_a c_a^T + F \qquad (2.2)$$

$$Y = \sum_{a=1}^{A} t_a c_a^T + G \qquad (2.3)$$

Where: u_a is the score vector relative to the a^{th} PC; c_a (or q_a in some notations) is the Y-weight vector relative to the a^{th} PC; F and G are the residual matrices.

Overall, the quantitative relationship between X and Y is given by X-weight vectors (w_a^*) and Y-weight vector (c_a). These weights are essential for the understanding of which X-variables are important in describing or predicting the Y-variables. Deviations from the X/Y correlation structure (outliers) and departures from linearity are normally uncovered by u/t-type plots (Eriksson *et al.*, 1999).

For both PCA and PLS models, the process variable data set (X) can be broken into meaningful blocks (X1, X2,...); each block may contain data from a single process unit or initial conditions such as raw material quality information. This approach is referred to as multiblock PCA or PLS (MB-PCA or MB-PLS) (Kourti and MacGregor, 1996 and Westerhuis *et al.*, 1998).

Number of Principal Components Selection

Several methods are available for choosing the number of principal components (A) that should be used to parsimoniously describe the original data set (Jackson, 1991). According to Kourti *et al.* (1995), cross-validation is the perhaps the most reliable of these methods; a detailed description of this technique is given by Wold (1978) and Nomikos and MacGregor (1995a). In summary, this method consists of successively keeping portions of the data out of the model (developed with a certain number of PCs) and then using it to predict the omitted data. The objective is to determine when the addition of another PC does not improve the model's predictive power (i.e. find the value of A). This purpose can be achieved using various statistical procedures or criteria, some of which are described in Table 2.1 using PCA as a reference. For PLS models these calculations are normally performed based on the predictability of Y.

Table 2.1 Description of various statistical tests for the cross-validation procedure.

Criteria name	**Relations used**	**Condition used in determining when no more PCs should be added to the model**
Minimum Press	$PRESS_a = \sum (x_{ik} - \hat{x}_{ik})^2$ Where: PRESS is the predictive residual sum of squares corresponding to each PC (a)	Minimum value of PRESS is achieved
R (Wold, 1978)	$R = PRESS_a / RSS_{a-1}$ Where: RSS_{a-1} is the residual sum of squares after a-1 PCs	R becomes greater than 1
W (Krazanowski, 1983)	$W = \frac{(PRESS_{a-1} - PRESS_a / D_m)}{(PRESS_a / D_r)}$ Where: D_m and D_r are the degrees of freedom required to fit and remaining after fitting the a^{th} component, respectively	W becomes greater than 1

With relation to the cross-validation issue, Nomikos and MacGregor (1995a) emphasize that there is no sound statistical test for this procedure and thus, the number of PCs needed in a latent variable model should be based on the overall picture that these criteria give. Additionally, Eastman and Kraznowski (1982) state that the decision of how many latent variables should be retained is also dependent on the purpose of the model (i.e. troubleshooting or prediction).

Model Building

Once the number of PCs to be used in a model has been established, their sequential calculation can be performed using the NIPALS algorithm (Geladi and Kowalski, 1986 and Wold *et al.*, 1987). This iterative algorithm can also be modified to

handle missing data that might be present in the original data set. An overview on the use of latent variable methods for missing data estimation purposes is given in Muteki *et al.*, 2005.

Model performance can be evaluated by simultaneously considering the relative amount of variation that is explained (R^2) and predicted (Q^2) by such model (Eriksson *et al.*, 1999):

$$R^2 = 1 - RSS / SSX \qquad (2.4)$$

$$Q^2 = 1 - PRESS / SSX \qquad (2.5)$$

Where: SSX is the total variation in the X-matrix remaining after mean centering.

When PLS is considered, R^2 and Q^2 values are normally calculated based on the Y data. It is also possible to calculate the explained variation of a single variable, be it a predictor or a response.

Process Data Analysis Using a Reference Latent Variable Model

Within multivariate SPC methods, a set of "in-control" or nominal data, subjected only to common cause variations, is used to build a reference model. This model can then used to either troubleshoot or monitor new data. New observations can be projected onto the latent variable plane defined by the reference model:

$$t_{a,new} = p_a^T x_{new} \qquad (2.6)$$

Projections on the X-model ($t_{a,new}$) can be entered into the t/u PLS inner relation to calculate $u_{a,new}$ (Eriksson *et al.*, 1999).

The location of an observation on the LV plane is given by it's score value or, alternatively, the Hotelling's T^2 equivalent, and the squared perpendicular distance of the observation from the plane (residual) is given by the squared prediction error (SPE) or Q-statistics (Kourti and MacGregor, 1995 and MacGregor, 2003):

$$SPE_i = \sum (x_{ik} - \hat{x}_{ik})^2 \qquad (2.7)$$

$$T^2 = \sum_{a=1}^{A} \frac{t_a^2}{s_a^2} \qquad (2.8)$$

Where s_a^2 is the variance of the a^{th} score.

Control limits can be applied to the T^2 and SPE values or charts (if on-line monitoring is intended). A description of these calculations is given by Nomikos and Macgregor (1995a). Violation of these control limits by the new data indicates a deviation from the nominal or modeled process conditions. It is normally expected that changes in relationships between variables, such as that caused by sensor failure, is detected by the SPE statistic, while changes in operating conditions, such as a grade change, are detected by the Hotelling's T^2 chart (Lennox *et al.*, 2000).

A set of statistical tools have also been developed to aid in the determination of an assignable cause for any deviations from normality (Kourti and MacGregor, 1996). This is important so that appropriate actions can be taken to, either compensate for the fault in real-time, or to avoid future occurrences. The most widely used set of diagnostic tools are based on calculating the contribution that each variable has on individual scores or the residual space (MacGregor *et al.*, 1994 and MacGregor, 2003):

$$cont_{a,j} \; to \; SPE = x_{ij} - \hat{x}_{ij} \qquad (2.9)$$

$$cont_{a,j} \; to \; \Delta t_a = \Delta x_j p_{a,j} \qquad (2.10)$$

Where: x_j is process variable measurement j; $p_{a,j}$ is the loading vector associated with the a^{th} principal component and the j^{th} variable;

2.2 MPCA and MPLS

Batch processes are finite with respect to duration. Batch process data is thus characterized by being three-dimensional in nature: J process variables are repeatedly measured throughout K time intervals for I batch runs. Figure 2.1 provides a visual representation of this X (I x J x K) matrix.

Multiway principal component analysis (MPCA) and multiway partial least squares (MPLS) are equivalent to performing PCA or PLS on a two-dimensional matrix obtained by unfolding and rearranging the original three-way matrix (Nomikos and MacGregor, 1994). There are six possible ways of performing this task; the choice of which one to use depends on the type of variation one wants to analyze in the data (Nomikos and MacGregor, 1995a). Nomikos and MacGregor (1992, 1994, 1995a and 1995b) introduced a batch-wise unfolding technique, which results in an X (I x JK) matrix, while Wold *et al.* (1998) proposed a variable-wise unfolding technique, which results in a X (KI x J) matrix.

For analyzing and monitoring of batch processes, batch-wise unfolding (arranging of vertical I x J slices side by side to the right, as shown in Figure 2.1) is considered as being the most significant method of matrix rearrangement since it allows for the modeling of the deviations of each batch from the mean trajectory (Nomikos and MacGregor, 1995a, Kourti et al., 1995 and Undey and Cinar, 2002). When combined with mean centering, this technique removes the main non-linear and dynamic components in the data, allowing for the use of linear modeling techniques. Due to these advantages, batch-wise unfolding will be performed throughout this work.

Subsequent to unfolding, alignment (described in section 2.3), mean centering and scaling are performed on the original batch data set. PCA or PLS methods can then be applied to this unfolded and pre-treated data matrix.

The main variations in batch trajectories captured by the latent variable model can be assessed in different manners due to the time varying nature of these observations (Garcia-Munoz, 2004). For the reference data set, each element of the score vector (t_a)

corresponds to a single batch and depicts the overall variability of this batch, throughout it's whole duration, with respect to other batches in the data set. Loading matrices from the MPCA analysis indicate how the variable measurements deviate from their average trajectories under normal operation throughout the whole batch run (Nomikos and MacGregor, 1995a). In other words, variables with a higher absolute loading value have a higher contribution to that particular latent variable at that point in time.

It is also possible to calculate overall values of SPE and T^2 upon completion of each batch (i) (Nomikos and MacGregor, 1995a and Undey and Cinar, 2002):

$$SPE_i = e_i e_i^T = \sum_{c=1}^{KJ} E(i,c)^2 \qquad (2.11)$$

$$T^2 = \frac{t_a^T S^{-1} t_a I}{(I^2 - 2)} \qquad (2.12)$$

Where: S is the (A x A) estimated score matrix.

In order to calculate SPE and T^2 values throughout the batch, the following equations can be implemented for each observation (i) at each time interval (k) (Nomikos and MacGregor, 1995a and Undey and Cinar, 2002):

$$SPE_{ik} = \sum (x_{ijk} - \hat{x}_{ijk})^2 \qquad (2.13)$$

$$T_{ik}^2 = t_{ika}^T S^{-1} t_{ika} \frac{I(I-A)}{A(I^2-2)} \qquad (2.14)$$

Equations 2.11 through 2.14 consider that the batch for which they were calculated has reached completion, in other words, all K sample points are known. When on-line monitoring of new batches is intended, this is not the case.

In principal, for on-line monitoring of batch processes using LV techniques, K different reference MPCA models are needed to estimate the scores and residuals for each current time sample k. This approach is however very computationally intensive and several techniques have been proposed to fill in the missing values relative to the future observations in X_{new} in order to allow for the construction of a single model (Nomikos

and MacGregor, 1995a): i) substituting all missing future values for zero and thus assuming that the batch will operate normally and not deviate from it's mean trajectory; ii) setting all future deviations from the mean trajectory equal to that at the last measured sample point (k); iii) treating all future values as missing within the MPCA model. Garcia-Munoz *et al.* (2004) shows that the missing data option (iii) yields superior performance relative to all others and thus this will be the method of choice throughout this work.

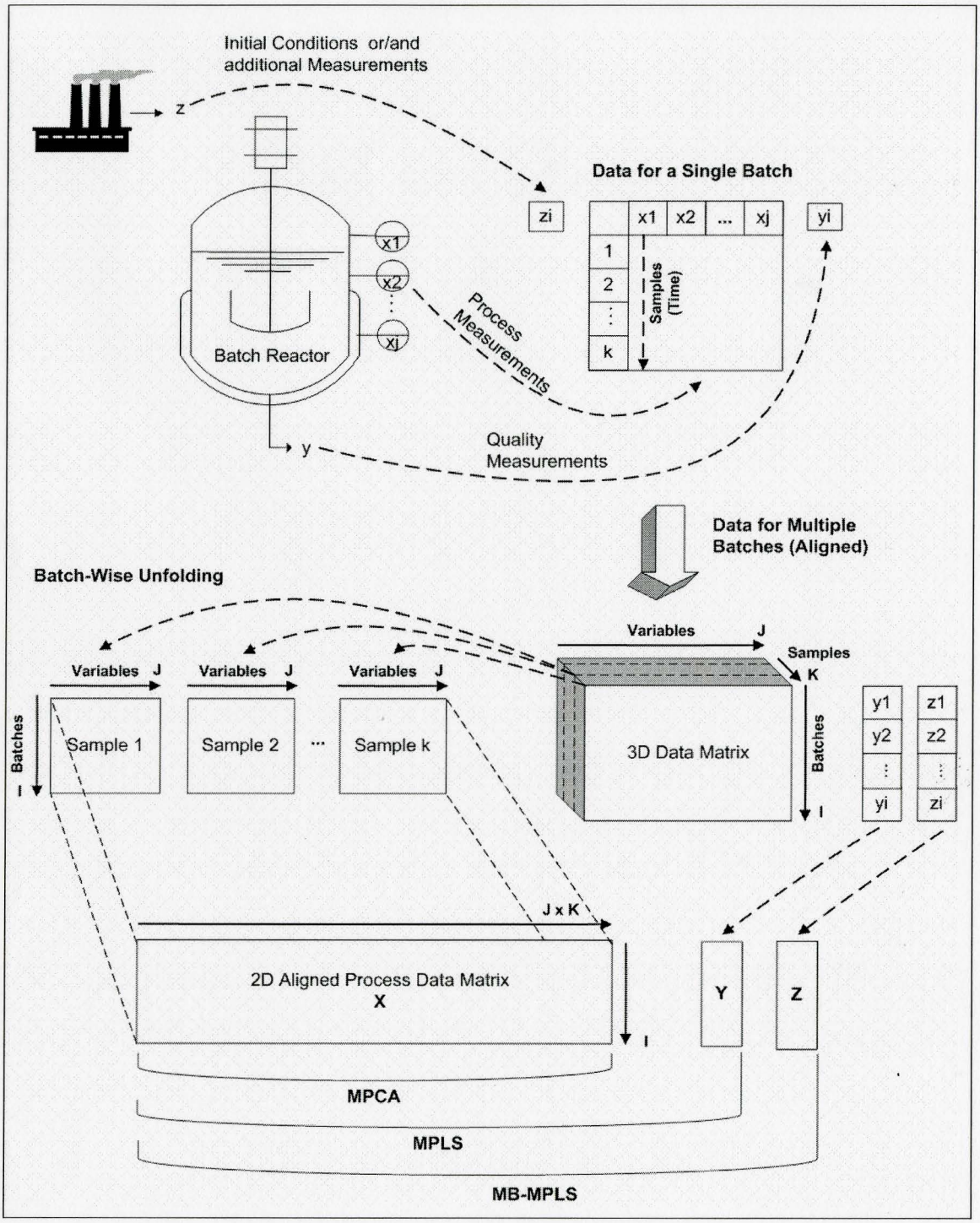

Figure 2.1 Nature of batch data: collection, batch-wise unfolding and latent variable model building.

2.3 Batch Data Alignment

One of the assumptions of MPCA and MPLS is that all batches have the same number of sample points (equal duration) and be synchronized or aligned in terms of variable trajectories (Nomikos and MacGregor, 1995b). Batches with equal number of sample points are essential due to matrix calculation rules and thus equalization of the batch lengths is mandatory prior to batch-wise unfolding. Additionally, alignment of the batch data is necessary to reduce extraneous variation in the variable trajectories. Models built with unsynchronized data will not provide as precise statistical results relative to fault detection as those built with synchronized data and will thus exhibit larger Type II error probabilities (Kassidas, MacGregor and Taylor, 1998).

Such conditions or assumptions are, however, not inherently true of batch data. Several factors can contribute to variability in total batch duration and/or the duration of it's various stages: seasonal differences in heating/cooling capacity, shared utilities, impurities in raw materials, non-automated steps which can be performed at the discretion of the operators, among others (Kassidas, MacGregor and Taylor, 1998).

The objectives of batch alignment or synchronization are thus to establish common start points at different phases of the run and to match variable trajectory shapes (Kourti, 2003). Additionally, it is necessary to ensure that all batches have the same number of sample points. Techniques proposed in the literature are:

- Trimming of the data set based on the batch of shortest duration (Marjanovic *et al.*, 2006).
- Augmenting the absent part of short duration batches with missing data (Kourti, 2003) or scaled deviations notes at the end of the batch run (Lakshminarayanan *et al.*, 1996).
- Crude linear interpolation over the entire batch time (Westerhuis *et al.*, 1999).
- Usage of an indicator variable with the objective of trajectory re-sampling which permits alignment (Nomikos and MacGregor, 1995a).

- Dynamic time warping (Kassidas et al., 1998).
- Usage of discrete events to determine the transition time between batch stages.

Trimming and data augmentation methods are easy to apply but, while they guarantee that all batches have the same number of sample points, these techniques (when applied by themselves) do not contribute towards trajectory matching. Satisfactory results in terms of alignment will be achieved only if the batches have variable time duration but overlapping trajectories in the common time part. In practice, this is a very restrictive condition (Westerhuis *et al.*, 1999, Kourti, 2003 and Kassidas *et al.*, 1998). Additionally, by trimming the data set based on the batch of shortest duration, all information relative to the end of the longer batches is lost. When data augmentation methods are used, restrictions exist relative to the percentage of missing data that can be handled by missing data algorithms.

Crude linear interpolation over the entire batch time results in a linear compression or expansion of each batch over it's entire duration (Westerhuis *et al.*, 1999). Thus, once again, while all batches treated in this manner will have the same number of sample points, their trajectories will only match if the increase or decrease in total batch time can be attributed to a proportional increase or decrease of the time spent on each production stage.

The use of an indicator variable for batch data synchronization was initially suggested by Nomikos and MacGregor (1995b) and has been successfully applied by Garcia-Munoz *et. al* (2003), Kourti *et. al* (1996), Neogi and Schlags (1998), among others. In order for this technique to be applicable, a monotonically increasing variable with fixed initial and end points must exist (Garcia-Munoz *et al.*, 2003). Such variable is what is being referred to as an indicator variable. Common examples of indicator variables are: conversion, cumulative weight of a key reagent, temperature ramps, among others. When a single indicator variable does not exist for the whole batch duration, different indicator variables can be selected for each stage of the batch (Garcia-Munoz *et al.*, 2003). Alignment using an indicator variable can be achieved by simply re-sampling

all other variable trajectories at pre-specified intervals of the indicator variable. This linear interpolation method is capable of guaranteeing both matching number of sample points and variable trajectories for all batches.

An alternative to the indicator variable method is dynamic time warping (DTW). Such method, initially introduced in the area of speech recognition, is capable of aligning batch data by translating, expanding or contracting certain localized segments within a batch with the objective of minimizing the distance between trajectories (Kassidas et al., 1998).

When the initial and final points of each batch stage are known, it is possible to re-sample all variables within this interval with respect to time. Identification of batch stages can be done through the use of discrete events within a batch such as: charging of reagents, heating, cooling and product discharge. Kaisha and Moore (2001) proposed a mathematical filter with the purpose of identifying such events from the batch data. Another procedure to help in phase identification is to take the derivative of a critical variable and verifying any changes in it's sign. Occasionally, such stages are recorded as an added variable in automated systems and thus a mathematical filter is not needed.

With all the alignment methods presented it is possible to include a variable expressing batch evolution (i.e. cumulative batch time) by augmenting the original process data matrix with this information (Westerhuis et al., 1999), as shown in the following section.

2.4 Batch Data Augmentation

When additional information such as a batch progression indicator, stochastic variables, data from other units, among others, are thought to contribute to the diagnostics capabilities of the model, the original data matrix (X) used for PCA or PLS can be augmented with these new variables (X_C) in the following manner (Yoon and MacGregor, 2001):

$$X_{Aug} = [X \mid X_C] \qquad (2.15)$$

Chapter 3

Troubleshooting of an Industrial Batch Process using Multiway Latent Variable Methods

The purpose of this chapter is to illustrate how the multivariate latent variable methodology proposed by Nomikos and MacGregor (1994, 1995a and 1995b) and Kourti et al. (1995) can be utilized, together with data pre-processing, in order to analyze and troubleshoot an industrial batch process.

More specifically, multiway principal component analysis (MPCA), multi-block multiway partial least squares (MB-MPLS) and batch alignment methods, were successfully used to troubleshoot a specialty chemical production unit in order to identify optimal process conditions with respect to quality. Additionally, approaches to data laundering of time-varying batch process variables were proposed.

3.1 Process Description

The process under study is composed of a reactor, which operates in semi-batch mode, a centrifuge and a storage tank. This unit is used to synthesize an organic specialty chemical referred to, in the present work, as LX. This production system is a critical step in a multi-step process and as such, LX is an intermediate product. A diagram of the LX production unit along with upstream and downstream process indications is shown in Figure 3.1.

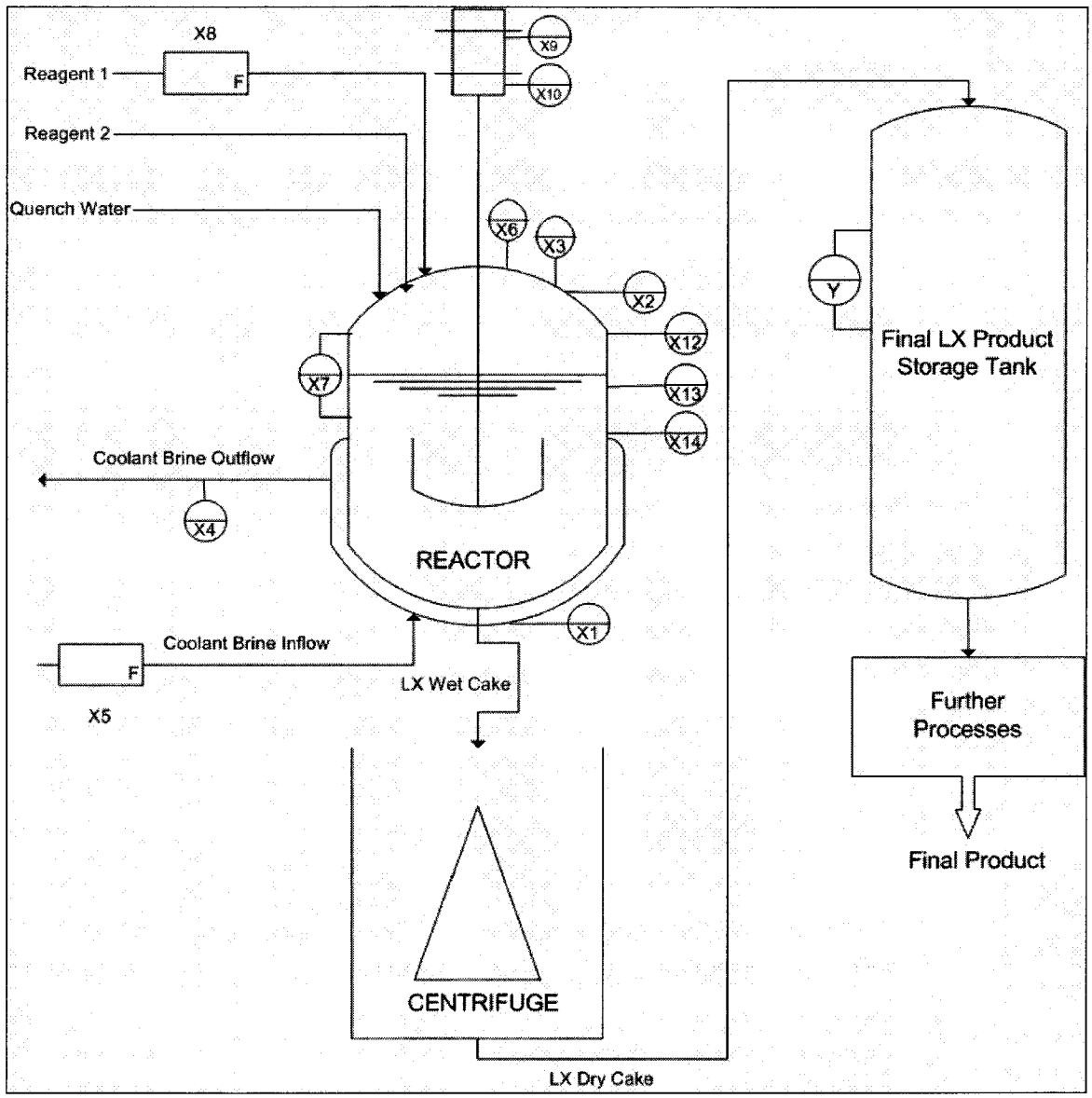

Figure 3.1 LX production unit scheme.

Recorded process variables for the LX batch process are listed bellow:

- X1 – Total Reactor Weight;
- X2, X12, X13 and X14 – Bulk LX temperatures in the reactor (X2 is the controlled variable);
- X3 – Head space temperature within the reactor;

- X4 – Coolant (brine) temperature at the exit of the reactor jacket;
- X5 – Coolant flow rate (manipulated variable in bulk temperature control);
- X6 – Surface temperature within the reactor;
- X7 – Liquid level within the reactor;
- X8 – Feed rate of critical reagent (reagent 2);
- X9 – Agitator Speed;
- X10 – Agitator Power;
- Y – Liquid level in the final product hold tank;

LX synthesis occurs within the reactor and follows a production recipe that is divided into four distinct phases. These phases are defined by the occurrence of a determined sequence of critical process steps or events, described as follows:

- Phase 1 – Addition of the first reagent (reagent 1) followed by initialization of agitation (which will be kept constant until the end of phase 4). Addition of reagent 1 is not directly measured; however, it can be inferred by an upward ramp in reactor weight (X1), as shown in Figure 3.2.
- Phase 2 – Addition of the second reagent (reagent 2) and control of the bulk temperature at a pre-defined, constant, set-point value. Reagent 2 is considered critical and yield determining and it's addition is marked by a second ramp in reactor weight (Figure 3.2) and recorded by variable X8.
- Phase 3 – Temperature rise to a new set-point value and addition of quench water (for dilution purposes). The increase in bulk temperature during this stage is due to an exothermic reaction between reagents 1 and 2.
- Phase 4 – Discharge of material to centrifuge. This is done in two steps, as can be seen by the two sudden drops in total reactor weight (X1) in Figure 3.2. The operators make the decision of when to discharge; delays maybe due to problems in the system (pump failures, equipment unavailability, among others). It was indicated by industrial personnel that if the diluted LX is held

in the reactor too long before discharge (1 or 2 hours) it becomes a bad intermediate material, due to degradation, and thus must be discarded.

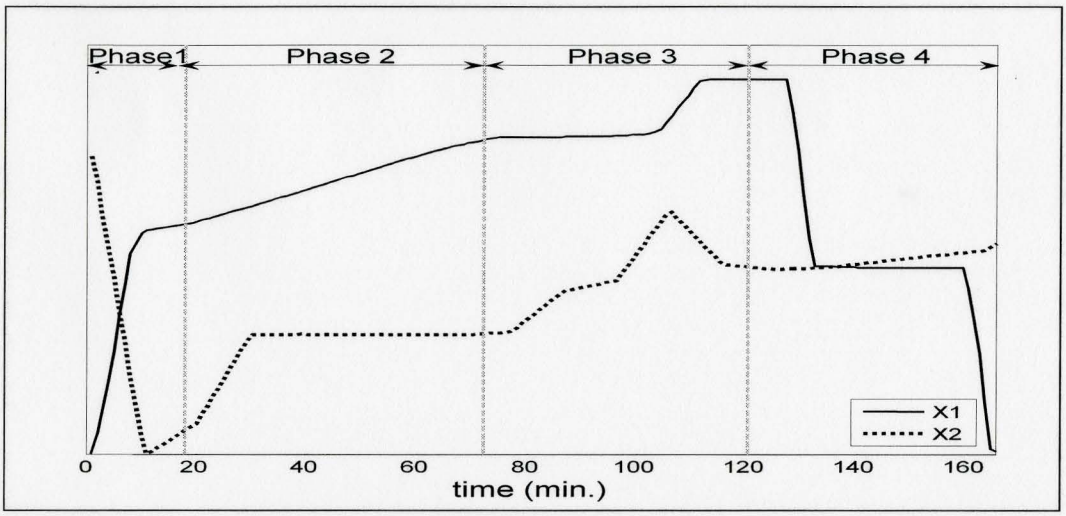

Figure 3.2 Reactor weight (X1) and bulk temperature (X2) trajectories during LX production phases.

After being discharged from the reactor, the synthesized material is centrifuged. No process measurements are taken during this production stage.

This product (LX) is then transferred to a hold tank, where it's total volume is indirectly measured by a tank level indication. This is the quality measurement (Y) for that particular batch run.

3.2 Project Objective

The present work aims at identifying optimal process conditions, with respect to final product quality, for the production unit under study. In practice, this translates to maximizing final LX liquid volume (Y) obtained per batch and minimizing it's variability throughout successive batches.

While no numerical target was set for such objective, it was indicated by industrial personnel that any incremental raise in Y values would lead to significant cost savings.

3.3 Data Set Description

The historical data set made available for the present study is composed of 58 LX batch runs. Each batch run contains thirteen process variables (X1 thought X10 and X12 through X14), one quality variable (Y) and batch tag discriminations. All of these measurements were registered by the digital control system (DCS) at every minute. Time consumption throughout each batch is inferred from this knowledge.

Batch tags are variables that give numerical indications that a discrete event or process step has occurred. There are 18 possible batch tags in all and these are executed in accordance with the process automation system. The operators are, however, able to overwrite this system by manual control.

LX production is carried out in monthly campaigns. Batch identification numbers are coded so that the first digit indicates the month of the campaign (i.e. batch 1051 was produced in January). Only data referring to the months of January, March and May was collected. This information is relevant in the subsequent discussions.

3.4 Data Pre-Treatment

The steps taken in order to treat the LX data set, prior to the multivariate analysis, are described in the sub-sections that follow.

3.4.1 Data Visualization

The first step taken towards data analysis was to plot all variables, for all batches, against time and visualize process behavior. The result of such preliminary analysis was

the exclusion of a batch due to extreme abnormal behavior of most of it's variables; in this case, batch beginning and end points could not be identified.

3.4.2 Alignment of the Batch Data

Preliminary data analysis showed that the batch runs under study have different durations and un-matched trajectories, thus indicating the need for alignment. As previously discussed in Chapter 2, there are various ways to align batch trajectories and their applicability and performance are case dependent.

Visual inspection of various LX trajectories, for both shorter and longer batches, indicated that these do not overlap in common time parts (Figure 3.5, left). Thus, neither crude linear interpolation with relation to total batch time or data augmentation of the shorter batches, are appropriate alignment methodologies. Additionally, the data does not present any single or combined process variables that can be used as indicator variables for the complete batch run.

In order to satisfactorily align the different batches from the data set in question, the four pre-defined process phases were further divided into nine stages and cumulative time within each of these stages was treated as the indicator variable.

Since the initial and final time of each stage is different for each batch, it is necessary to attribute a progression measure to allow for re-sampling. Thus, a specific stage was set as being 0% complete at it's first original sample point (beginning of the stage) and 100% complete at it's last original sample point (end of the stage). Each stage (s) of each batch (i) was then re-sampled at time increments given by $\Delta Time_{is} = Time_{is}/(n-1)$, where n is the desired number of samples for that stage (Garcia-Munhoz *et al.*, 2003). For the current work, re-sampling was done using linear interpolation.

Identification of the stages to be used for LX data alignment was done through the recognition of discrete events that occurred throughout all batch runs. Since the occurrence of such events was registered in the historical data sets as an additional variable (batch tags), they were easily recognized and thus no mathematical filtering was

necessary, except in the identification of the discharge stages. For these last three stages (corresponding to phase 4) a second order derivative of total reactor weight (X1) with respect to time (T) was used to aid in the determination of beginning and end points of discharge in an automated manner. Such points show up as pronounced peaks or dips in plots of absolute values of d^2X1/dT^2 *versus* time (T) (Figure 3.3).

Various numerical techniques can be used to calculate derivatives and partial derivatives of batch trajectories with respect to their evolution index (Garcia-Munoz, 2004). Such techniques vary in complexity and robustness to noise. In this work the approach taken was simply to calculate the difference between two adjacent elements of X1 in time (single point derivative):

$$\left.\frac{\partial X_1}{\partial T}\right|_{t=k} = \frac{X_{1(k)} - X_{1(k-1)}}{T_{(k)} - T_{(k-1)}} \qquad (3.1)$$

$$\left.\frac{\partial^2 X_1}{\partial T^2}\right|_{t=k} = \left.\frac{\partial \dot{X}_1}{\partial T}\right|_{t=k} \qquad (3.2)$$

The new process stages defined for alignment are shown in Figure 3.4.

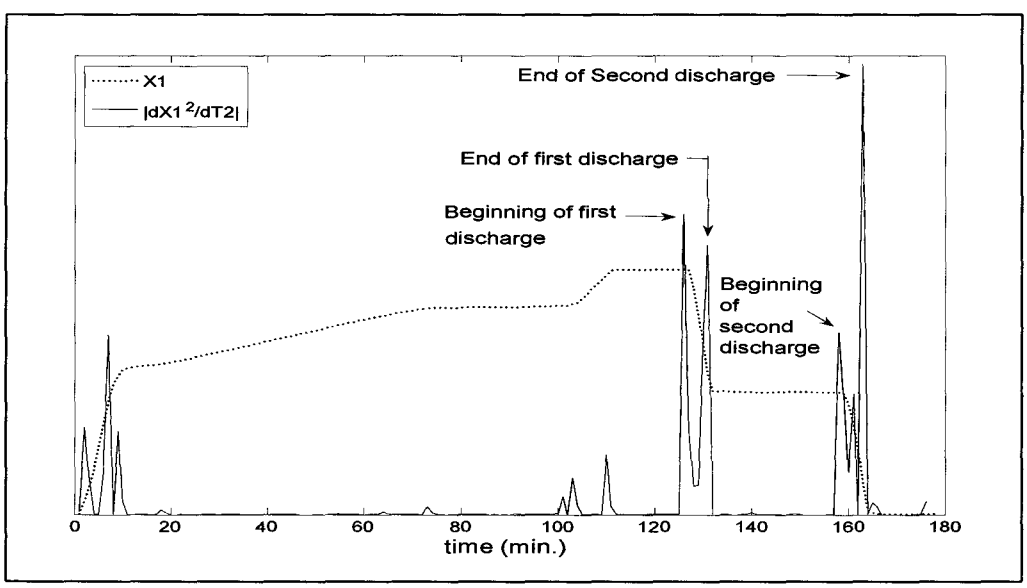

Figure 3.3 Plot of total reactor weight (X1) and absolute values of it's temporal second order derivative (d^2X1/dT^2) *versus* time for the identification of process stages 7-9.

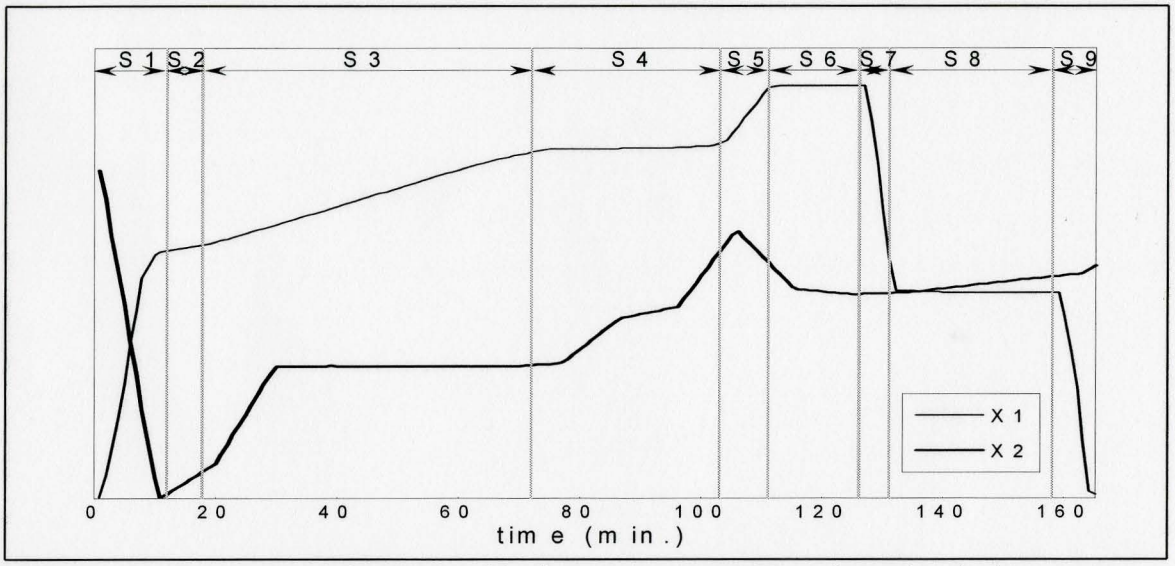

Figure 3.4 Reactor weight (X1) and bulk temperature (X2) trajectories during process stages (S) used for LX batch data alignment.

The number of observation within each alignment stage (n) used during re-sampling was determined by the average number of observations before alignment for those batches considered "normal" (Table 3.1). This is done to avoid errors caused by interpolation resulting from having too many or too few actual measurements between two interpolated values (Kourti, 2003).

Table 3.1 Vector of average number of observations within each alignment stage (n), prior to alignment.

Stages	1	2	3	4	5	6	7	8	9
Observations	13	4	56	30	8	25	5	28	5

Visual inspection of plots of total reactor weight (X1) and bulk temperature (X2) for all batches (Figure 3.5), indicates that the alignment methodology used was successful at ensuring that all batches have the same number of sample points (174) and synchronized trajectories.

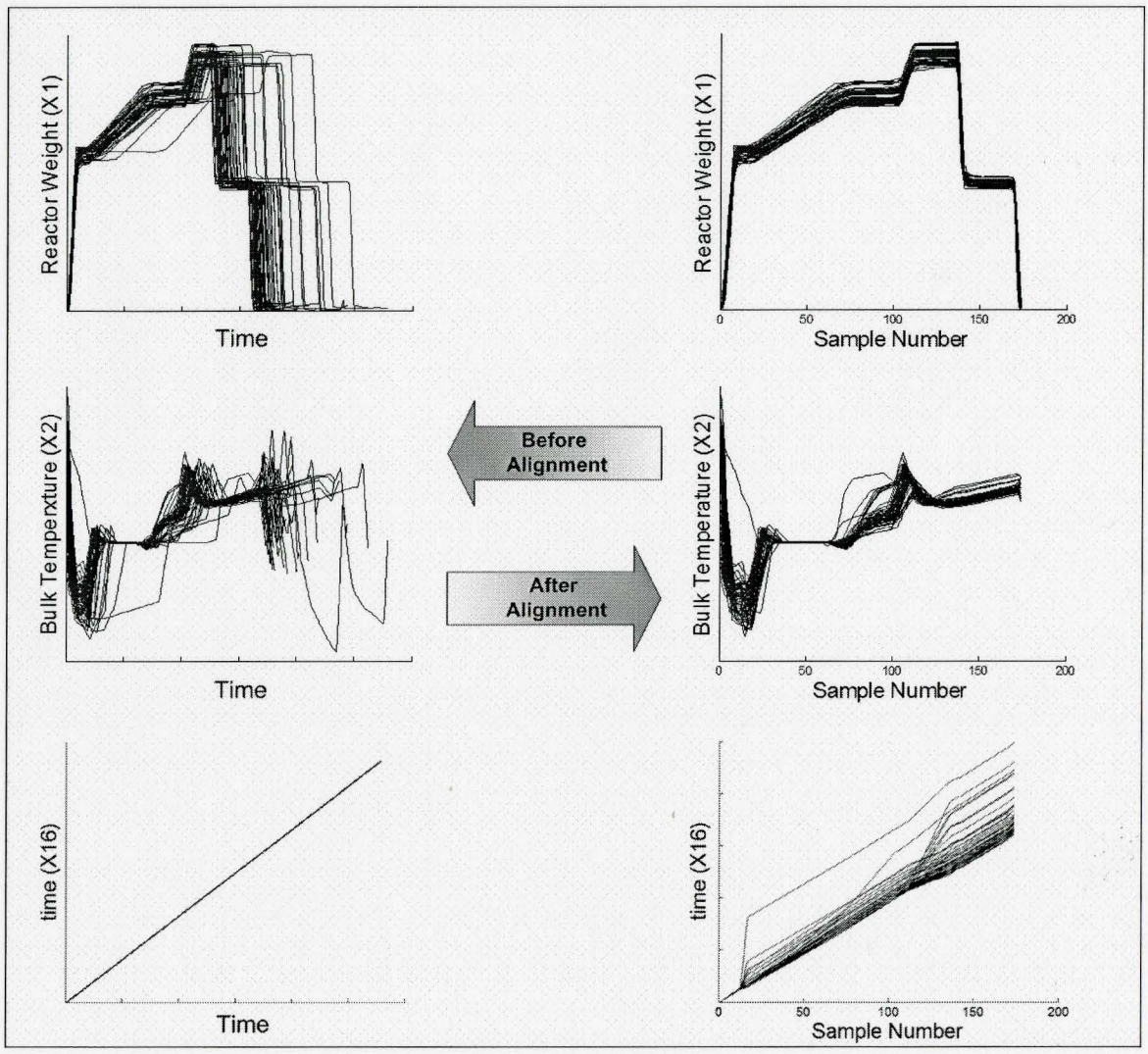

Figure 3.5 Plots of variables X1, X2 and X16 for all batches, before alignment (Left) and after alignment (Right).

3.4.3 Data Augmentation with Calculated Variables

Subsequent to batch alignment, better fault detection and identification is possible through the inclusion of cumulative time as an extra variable in the data set (Westerhuis *et al.*,1999, Garcia *et al.*, 2003 and Kourti, 2003). Thus, in the current case-study, time usage per batch was incorporated as the 16^{th} variable in the X matrix. Visual inspection

of Figure 3.5 indicates that, prior to alignment, time evolves linearly for all batches while, subsequent to this transformation, it is "distorted" and incorporates information regarding how each batch "evolves" through time.

With the intent of increasing data interpretability, new variables, expressing the cumulative sum of coolant and reagent 2 flow rates (X5 and X8), were also added to the X matrix as the 14^{th} and 15^{th} variables respectively. These new trajectories, referred to as X5s and X8s, rely on the use of previous knowledge in order to express in a more meaningful manner the total heat transfer from the reactor and the total amount of reagent 2 added throughout the batch progression.

Finally, based on knowledge gained from a preliminary MPCA study, the data set is augmented to contain an extra Z matrix allowing for an MB-MPLS analysis. This matrix will be better described in section 3.5.2 of this thesis.

3.4.4 Unfolding, Scaling and Mean Centering

The current work relied on the use of the software BatchSPC version 2.0 developed by the McMaster Advanced Control Consortium (McMaster University, Hamilton, Ontario, Canada). This program automatically unfolds the original data matrix in a batch-wise manner and auto-scales it (scales to unit-variance and mean-centers) in a column-wise manner.

It is also worth mentioning that, throughout the course of this Chapter, cross-validation was performed on the unfolded data matrix using Simca P+ version 11, developed by Umetrics. This software relies on R-statistics (described in Chapter 2) and a set of rules to determine the optimum number of principal components in a model.

3.5 Troubleshooting of the Batch Data

The main objective of this section is to build latent variable models using historical data from the LX production unit, to analyze the process and determine which variations in the process variables have the highest impact on final product quality.

The analysis is presented in three sections. In the first section (3.5.1), an initial assessment of the data is performed using MPCA. Based on knowledge gained from this preliminary study, a second section (3.5.2) shows how the data set is augmented to contain an extra Z matrix and a multiblock, multiway, PLS model is fitted and analyzed. The third section (3.5.3) is aimed at applying data laundering techniques to extract further quality-relevant information from the historical data.

Theoretical concepts relative do MPCA and MB-MPLS are described in Chapter 2 of this thesis.

3.5.1 Multiway Principal Component Analysis (MPCA)

Even though the primary objective of the current project is to determine which variations in process conditions are the most influential on the quality variable (Y), it is also important to understand the variability within the process variables (X) as a whole. This allows for gain in process knowledge and the identification of clusters of operating conditions. Such clusters are indicative of any significant changes or deterioration in the process during the data-collection period.

Detection and understanding of changes in operating conditions is of extreme importance since data based or inferential models are only valid for the set of process conditions for which they were identified (reference dataset). This is described in the literature as the fundamental assumption of comparable runs (Kourti *et al.*, 1995).

Processes that run based on campaigns are especially prone to presenting shifts in operating conditions since they are normally used in multi-product synthesis and are thus subjected to contamination, alterations in operating set points and even mechanical

configurations. If a process shift is detected it must be carefully studied. Data from periods prior to the occurrence of such changes may have to be excluded, since they no longer represents "normal" operating conditions.

With these objectives in mind, a MPCA model was built using a single X matrix containing all 57 pre-treated batches from the historical data set. Each batch contained all 13 original process variables and the 3 calculated variables.

Outlier Detection

Methods for calculating the optimal number of principal components (PCs) a model ought to have, such as cross-validation, should not be applied on a first analysis of the historical data. This is due to the fact that, most likely, the data contains some outliers that will affect the performance of such techniques, depending on their robustness.

Initial outlier identification was performed by building a model with the 2 principal components (PCs) that captured the most variation in the historical data set, and evaluating the resulting overall Hotelling's T^2 and SPE values for the individual batches. Batch 5121 was identified as a strong outlier due to the fact that it presented an overall Hotelling's T^2 value significantly above the 99% confidence interval. A model built after the exclusion of batch 5121 was "interrogated" to determine what caused this batch to have a projection to the score plots (t1 and t2) so different from all other batches. An overall contribution plot for the first score (t1) between this batch and averaged values from all other batches indicated that batch 5121 has an abnormal behavior with relation to variables X5s, X14 and time. This was confirmed through the inspection of the raw data plots and. Batch 5121was thus determined to be an outlier and excluded from the data set.

Interpretation of the MPCA model

A final MPCA model, containing data from the remaining 56 batches, was subsequently built. Such model uses 3 latent variables to capture 56% of the variance in the X matrix; 32% of this variance is captured by the first principal component alone.

Cross-validation initially showed that 8 PCs were optimal capture 78% of the variance in the process variables (Figure 3.6). Since the aim of the current section is to analyze the main variability in the data, only components that increased the fraction of the sum of squares explained (R^2) by more than 10% where kept.

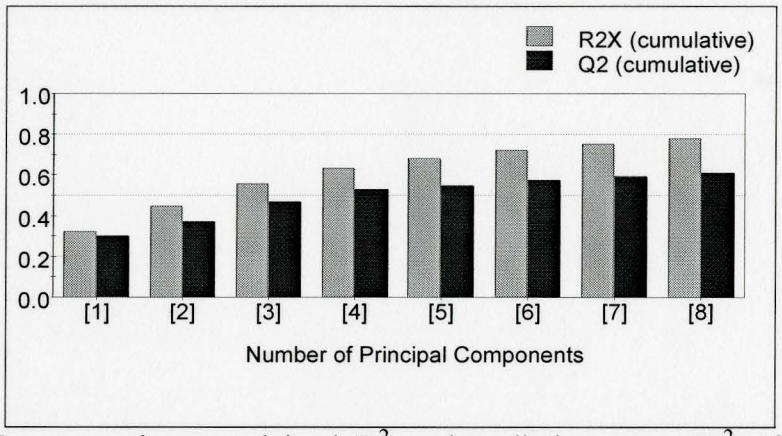

Figure 3.6 Percent variance explained (R^2) and predictive power (Q^2) of the models containing the number of PCs listed in the horizontal axis.

Figures 3.7 and 3.8 show the projections of all 56 batches onto the score planes (t1/t2 and t1/t3) of the final MPCA model. Visual inspection of such plots shows the existence of four obvious data clusters in the t1/t2 score space and three, less obvious ones, in the t1/t3 score space. A couple of things are readily apparent when examining these clusters: (i) each cluster is composed mainly of batches run in the same production campaign; (ii) it is probable that two production campaigns were carried out in January; this is supported by the fact that the batch data identification numbers differ only in the third digit for those batches pertaining to different clusters; (iii) clusters are separated mainly along the direction of the first score space (t1), however, both t1 and t2 or t3 are needed to see the separation.

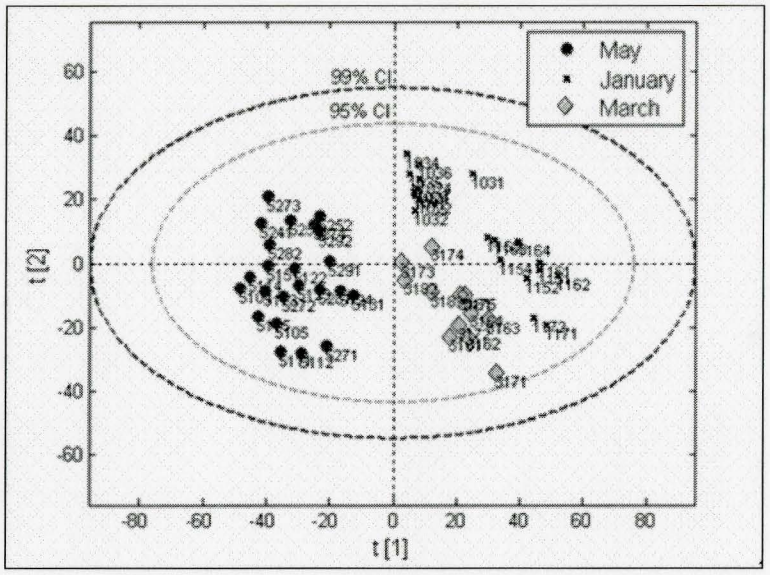

Figure 3.7 Score plot for t1/t2.

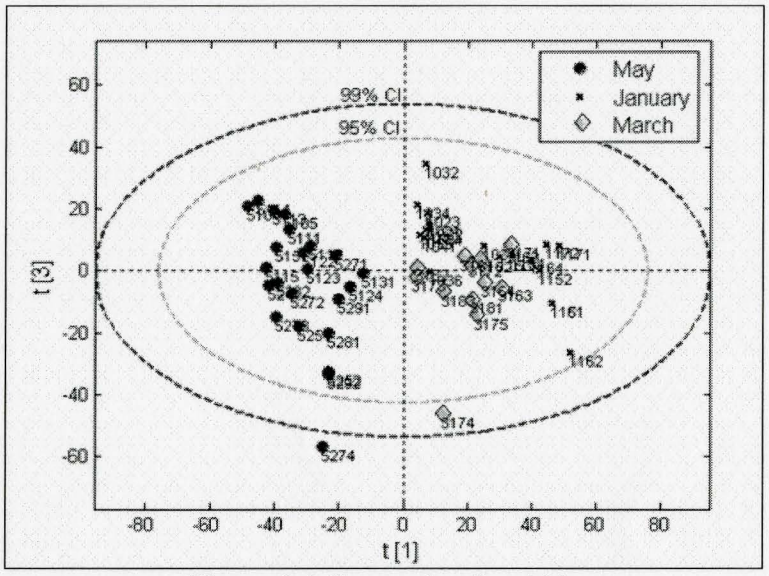

Figure 3.8 Score plot for t1/t3.

The existence of data clusters in the score plots indicates that batches run in the same production campaign have similar behavior in relation to their variable trajectories, while those that where produced at a different campaign, do not. It is thus necessary to further interpret the model to determine which variables are responsible for such disparity. However, due to the time varying nature of these observations, the main

variations in batch trajectories captured by the latent variable model can be assessed in different manners.

The loadings relative to the first PC of the MPCA model in question (Figure 3.9) point to the importance of variables X1, X3, X5 (and X5s), X6, X9, and X10 in explaining the variations in the t1 direction. Furthermore, X4 and time also where shown to contribute to this purpose, even if not as significantly. While X4, X9 and time are all positively correlated with each other, X1, X3, X5, X6 and X10 are all negatively correlated with these first variables but positively correlated among themselves. This means that higher than average values of X4, X9 will lead to an increase in t1 values and higher than average values of X1, X3, X5, X6 and X10 will lead to a decrease in t1.

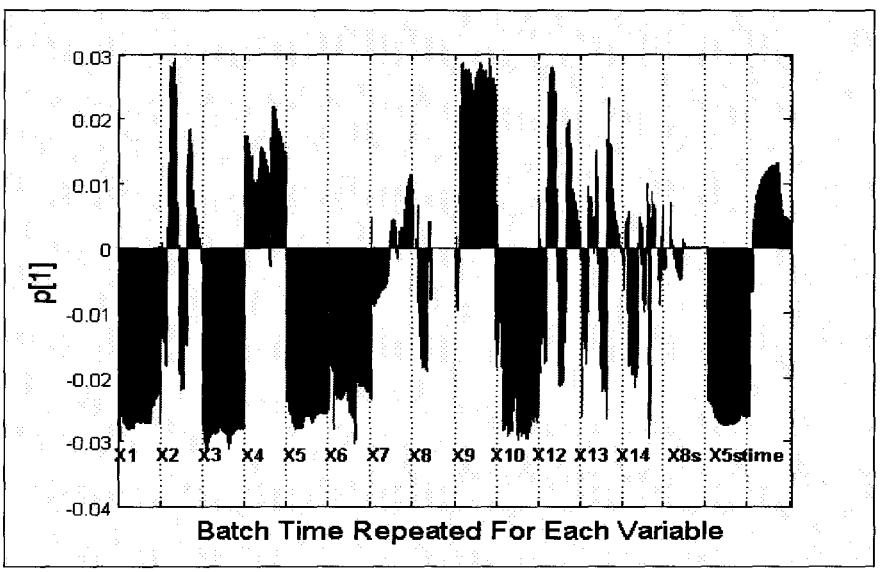

Figure 3.9 MPCA loading plot for the first principal component.

The final and most significant step taken in this troubleshooting exercise was the analysis of various contribution plots between batches belonging to different clusters in the t1 score space (Figure 3.10). The following conclusions were drawn from these graphs: i) batches produced in January (both clusters) have consistently lower values of X3, X5, X6 and X10, and higher values of X9 than both March (Figure 3.10, top left) and May (Figure 3.10top right); ii) batches produced in March have consistently lower values of X3, X5, X6 and X10 and higher values of X9 than May (Figure 3.10 bottom left); iii)

the main difference between the January clusters is that the batches with lower identifications numbers have consistently higher values of X3 and X6 and than those with higher identification numbers (Figure 3.10 bottom right); iv) X1 does not consistently vary between clusters.

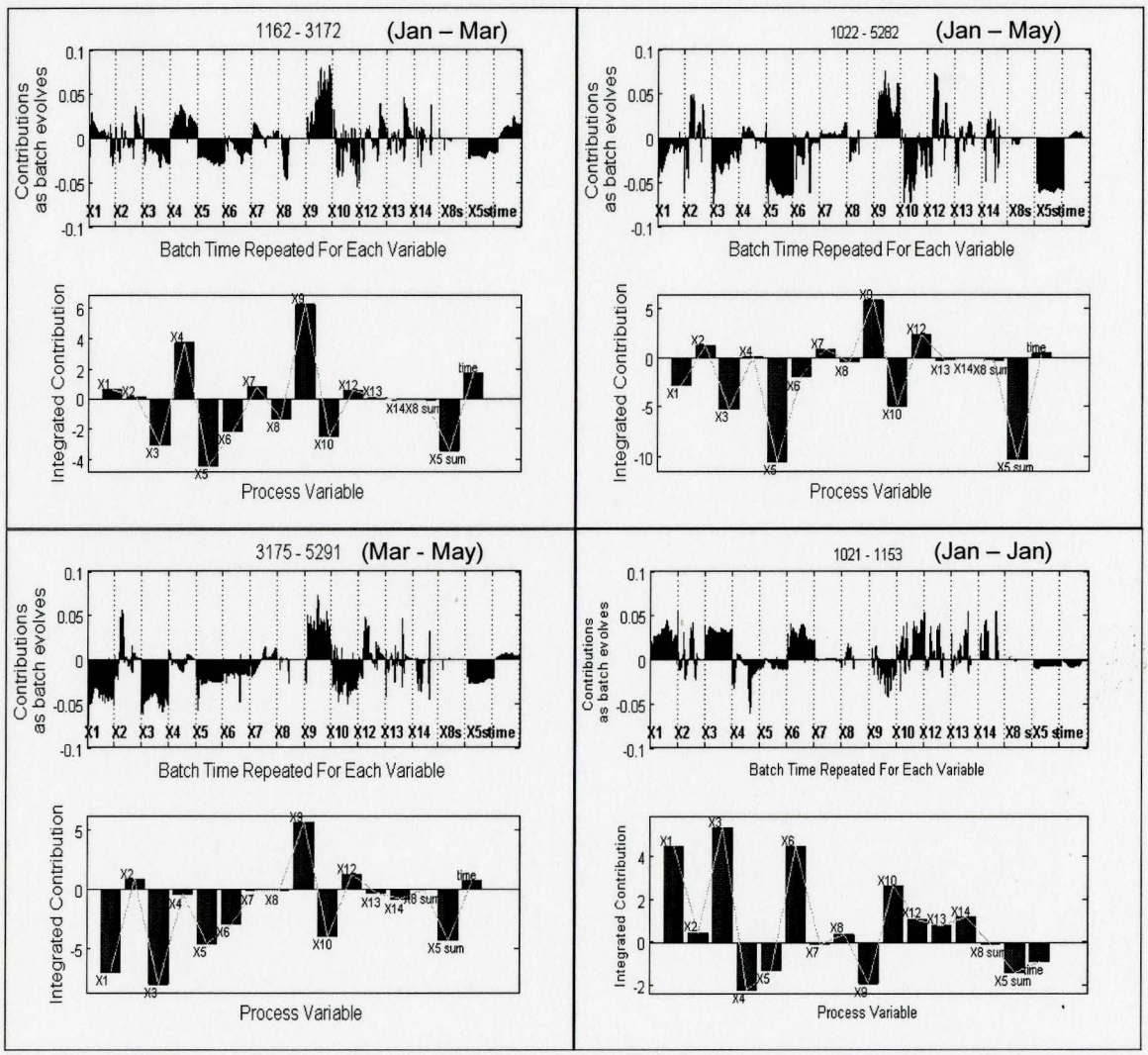

Figure 3.10 Contribution plots for the first score (t1) for all variables, over all time points, between batches: 1162 and 3171 (top left), 1022 and 5282 (top right), 3175 and 5291 (bottom left) and 1021 and 1153 (bottom right).

It is known that the LX product plant is located in a southern American state and thus subject to very hot summers and fairly cold winters. A crucial observation resulting

from the analysis of the contribution plots between clusters, is that most of the variables that were consistently lower in the earlier, and thus colder, months of the year are temperature related (X3, X5 and X6). This suggests that such variables are subjected to seasonal variation. Based on these observations, it is possible to explore potential scenarios for the LX process behavior depicted by the variables within this process:

- As mentioned in the previous paragraph, ambient temperatures in May are warmer than those in March, which are, in turn, warmer than those in January. This may cause the inflowing coolant temperature to also rise in the summer months, possibly due to heat transfer during transportation within the pipes or even to lower brine refrigeration performance. Since the brine coolant is used to control the product bulk temperature within the reactor (X2, X12 and X13), a larger brine flow rate (X5) is needed to compensate for higher temperatures in the in-flowing coolant. Such phenomenon can also cause the out-flowing brine temperature (X4) to increase.

  The effect of seasonal changes in cooling water temperatures and, consequently, in heat removal capabilities has been previously described by Kassidas *et al.* (1998).

  In hindsight, the seasonal effect on X5 can be seen univariately (Figure 3.11 left), however, such effect would be very difficult to pinpoint using this simpler approach.

- Higher values of surface (X6) and headspace (X3) temperatures during summer month can be justified in two possible ways: i) in case the temperature sensors in question are located close to the reactor wall, they may be influenced by the brine flow temperature within the reactor jacket and thus present higher temperature readings for the summer months; ii) it is known that the LX reaction takes place in a nitrogen rich atmosphere and that the N_2 is kept in exterior tanks with no temperature control. Thus, if X6 and X3 measurements are influenced by

the nitrogen temperatures within the tank, they will exhibit higher values in hotter production days.

- The inverse relation between agitator speed (X9) and power (X10) observed between seasonal clusters is on a first analysis contradictory since, fundamentally, such variables are directly proportional. However, the following hypothesis can be used to explain such phenomenon: i) both metallic and polymeric parts are known to expand when submitted to warmer temperatures, thus, summer months could present higher friction between the motor rotational parts leading to an increase in power consumption (X10) and decrease in speed (X9); ii) differences in temperature can accentuate the formation of impurities within the raw material which could lead to changes in product viscosity and thus increase agitator power consumption and decrease speed.
- As previously mentioned, increase in total reactor weight (X1) with increasing ambient temperatures is not fully proven due to lack of consistency. It maybe that more batches of the May cluster show higher values of X1 and thus this variable is confounded with the seasonal effect on temperature. This claim is again supported univariately in Figure 3.11(right).

It is important to note that the scenarios listed above are consistent with respect to the positive and negative correlations between variables described by the loadings plot. In this case, ambient temperature is a lurking variable and it plays a major role in affecting the relationship among X3, X4, X5, X6, X9 and X10.

Industrial personnel responsible for LX production indicated that the scenarios previously described to explain the patterns seen in the data, where very plausible. However, it should be noted that all causal statements are based on previous process knowledge; since they cannot be inferred form the multivariate models alone.

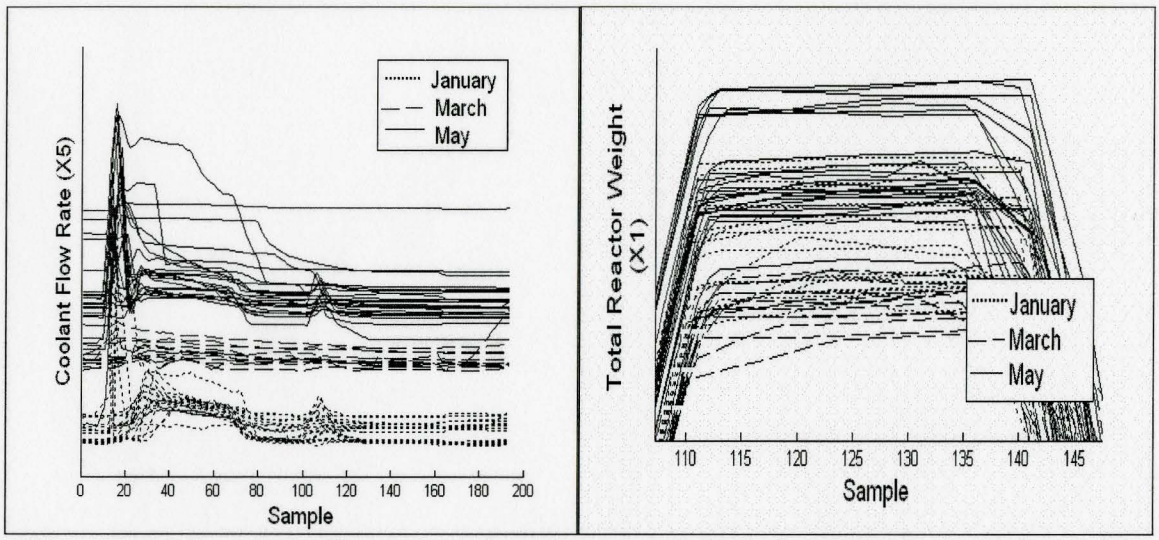

Figure 3.11 Seasonal effects on coolant flow rate (left) and total reactor weight (right).

In summary, the MPCA model allowed for a significant gain in process knowledge relative to the seasonality of the LX production unit. However, it is not entirely correct to state that there was a shift in process conditions between production campaigns, since the differences observed may repeat themselves annually. It is, nevertheless, necessary to keep such changes and the resulting confounding effects in mind when proceeding with the data analysis.

Additionally it was noted that changes in total reactor weight (X1) play a major role in the variability of LX production. However, due to the cumulative nature of this variable it is not possible to explicitly distinguish which reagents contribute to such deviations. This observation is important and measures are taken in the following section to further facilitate diagnostic procedures with relation to product quality. The same observation applies to process time usage.

3.5.2 Multi-block, Multiway, Partial Least Squares (MB-MPLS)

The objective of the current section is to determine which variations in the process variables of the LX production unit are most influential on the final product quality.

As previously mentioned, both total reactor weight (X1) and time usage are believed to play a major role in the current analysis. However, the cumulative nature of these variables makes their interpretation difficult. In order to overcome this problem, selected information from the original X1 and time vectors were added to a new block of data (Z). The information contained in the Z matrix is: total time necessary to complete stages 1 and 2 (tR1), 3 (tR2), 4 (tReact), 5 (tW), 6 (tplato) and 7 to 9 (tdisch) and total weight of reagents 1 (R1), 2 (R2) and water (W) added per batch.

In order to achieve the objective of this troubleshooting exercise and considering the nature of the data set, a multi-block, multiway PLS analysis was performed. This regression model is capable of simultaneously relating the final product quality data (Y) to the aligned process trajectory data (X) and the additional matrix (Z).

Outlier Detection

Outlier identification was performed through the inspection of a preliminary MB-MPLS model fit with 2 PCs (equal weights were given to the X and Z matrices). Batches 1151, 5121 and 5151 were identified as having very high distances from the model (significantly above the 99% confidence interval for the SPE) with relation to the Z-space, Y-space and X-space respectively.

Overall contribution plots for SPE between batches 5121 (X-space) and 1151 (Z-space) and an average from all other batches, again shows that batch 5121 has anomalous behavior with relation to variable X14, while batch 1151 spends a much higher than average time in stage 5 (tReact). Inspection of the raw data for batch 1151 indicated that the DCS system erroneously attributed the same batch tag for various stages. Both of these batches were excluded from the data set.

Batch 5151 presented an extreme distance from the model in the Z-space due to the fact that it has a much lower than average value of Y without having any deviations in the X-space that would cause this. This is an indication that there was probably some error in the collection of the quality data point for this batch and it was thus, also excluded from the data set.

Batch Classification

With relation to batch classification, no direct indication of whether a batch progressed in a desired manner with relation to it's process or quality variables was given. LX industrial personnel only stated that batches that yielded higher product volume (Y) were extremely desirable.

However, one of the fundamental assumptions of all inferential approaches is that of observable events (Kourti *et al.*, 1995). In other words, it is crucial to establish if the model is capable of discriminating between "good" and "bad" batches from the measurements that were collected.

Considering these statements, a decision was made to classify the batches within the LX data set based on their values of final hold tank level. Visual inspection of Figure 3.12 shows that, due to their high values of Y, batches 5104, 5105, 5111, 5112, 5113, 5114 and 5115 are, for the purposes of the current study, considered as being "very good" and batches 1031, 3161, 3171, 3172 and 5271 as being "very bad".

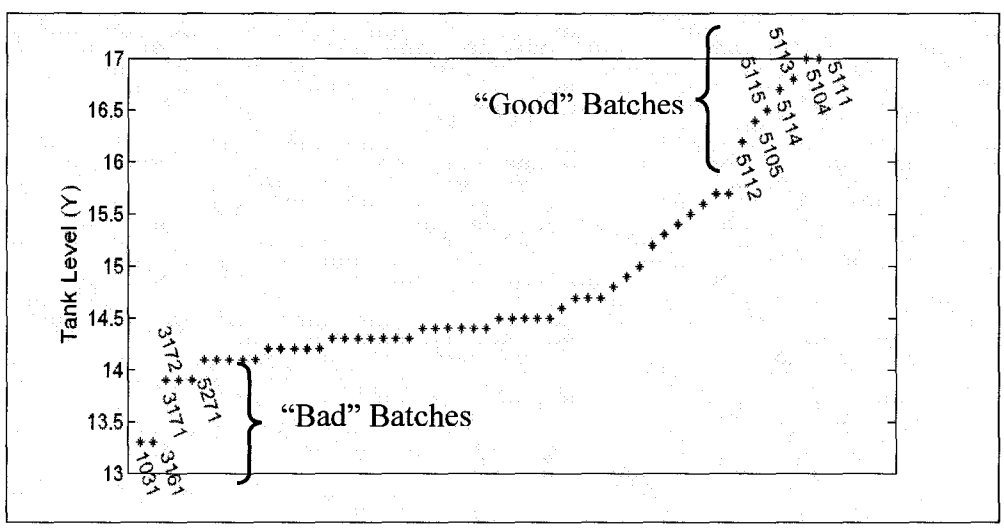

Figure 3.12 Batch classification based on values of the Y data.

Block Weight Selection

In order not to bias the multi-block PLS model, it is important to attribute the right amount of "importance" to the X and the Z matrices; this can be achieved by attributing weights to them (Westerhuis *et al.*, 1998). To determine these weights, preliminary analysis or knowledge of the system can be used. MacGregor *et al.* (1994) proposed a way of verifying that successful blocking has been achieved by comparing the predictions of Y achieved from the single and multi-block models for the same number of PCs. Essentially this is a trial-based approach; if the multi-block model provides lower predictability then the single-block PLS model, different weighting or blocking arrangement or an increase in the number of PCs included in the model, should be attempted.

As an initial guess, equal weights were applied to both the X and Z matrices and a X-Z-Y MPLS model was fitted. In order to verify the validity of this choice, the results for this multi-block model along with those for the single X-Y MPLS and Z-Y PLS models are summarized in Table 3.2.

Table 3.2 Summary of the results of various PLS models for block weighting.

Model	**X-Y MPLS**		**Z-Y PLS**		**X-Z-Y MPLS**	
Component	**R^2 Y (%)**	**Q^2 Y (%)**	**R^2 Y (%)**	**Q^2 Y (%)**	**R^2 Y (%)**	**Q^2 Y (%)**
1	53.9	49.2	75.7	71.9	67.1	62.7
2	83.8	79.2	80.7	74.5	83.9	78.6
3	91.2	85.0	82.2	73.2	90.9	84.8

Inspection of Table 3.2 shows that no significant predictive power is lost between the multi-block and the single models. Thus equal weights for the X and Z matrices are applied to the final multi-block model.

Additionally, Table 3.2 also shows that the sum of the captured variance (R^2) and the predictive power (Q^2) for the multi-block model is practically the same as that of each of the single blocks. This indicates that the X and Z blocks are not orthogonal from each

other and of equal importance in describing the final quality variability. The first observation is expected since the Z matrix contains a subset of the information contained in X. The second observation is, however, not expected. This means that the variables selected to compose Z, which are much fewer than the ones contained in the X matrix, are very influential in the final product quality.

MB-MPLS Model

A MB-MPLS model, containing data from the remaining 54 batches and having equal X and Z block weights, was subsequently built using 2 PCs. The amount of variance in the Y vector captured by this model was very high (84%), especially considering that no process measurements were taken within the centrifuge. This indicates that variations within the reactor are almost exclusively responsible for the final quality achieved.

The variance captured by the model in the X- space and Z-space were: R^2X = 38% and R^2Z = 27%, respectively. Additionally, no departures from linearity between X, Z and Y were verified.

Multi-block PLS models have two sets of scores: i) the super-scores (TT), which take all blocks into account in explaining Y; ii) the block scores, which will take one block (X or Z) into account at a time to explain Y (Garcia-Munoz, 2003).

Visual inspection of the TT1/TT2 super-score plot (Figure 3.13) shows a clear clustering of batches 5103, 5105, 5111, 5112, 5113, 5114 and 5115. Also, clustering of the batches pertaining to the same production campaign is verified, but to a much smaller extent.

With relation to quality, the batches pertaining to the cluster singled out in the super-score plot are those that were considered as the "good", or in this case "best", batches, due to their high values of Y. These batches have high TT1 values. Inspections of the batches with the lowest TT1 values shows that these are, for the most part, those that were considered as the "bad" batches due to their low values of Y. This proves that

the system is "observable" and thus that the process data collected is able to discriminate between "good" and "bad" batches.

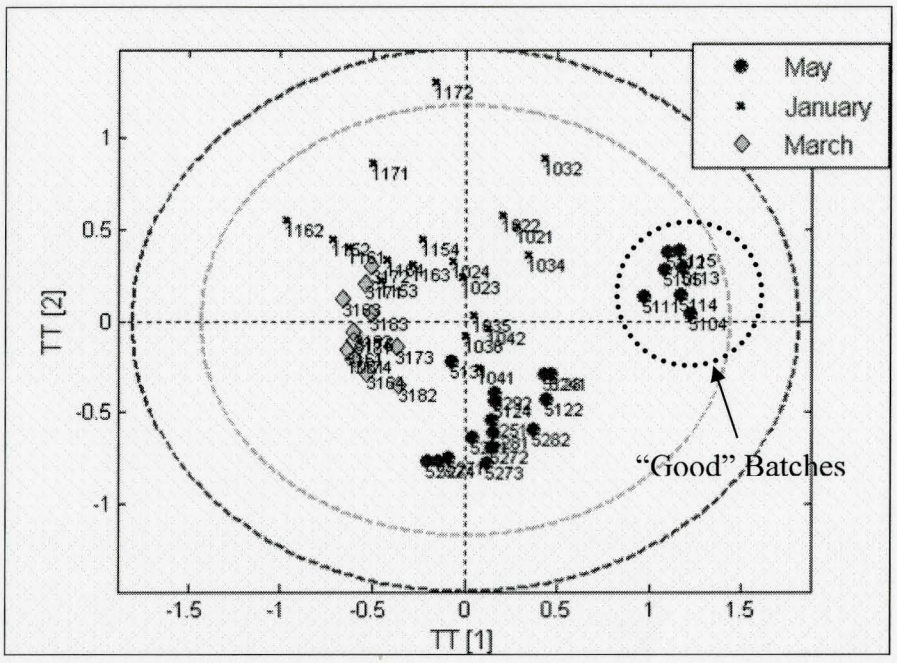

Figure 3.13 Super-score plots TT1/TT2, clustering of the "good" batches is identified.

In order to isolate the cause of the clustering of all the "good" batches, an absolute contribution plot for the first PC between one of these "good" batches (5115) and another batch (5123), with lower Y value, also produced in May was analyzed (Figure 3.14). This plot points to higher values of reagent 2 added to batch 5115 as being the main contributor to the difference in quality observed. This phenomenon was consistent for all batches in the cluster depicted in Figure 3.13: for all "good" batches the total weight of reagent 2 added to the reactor was significantly larger than for the remaining batches.

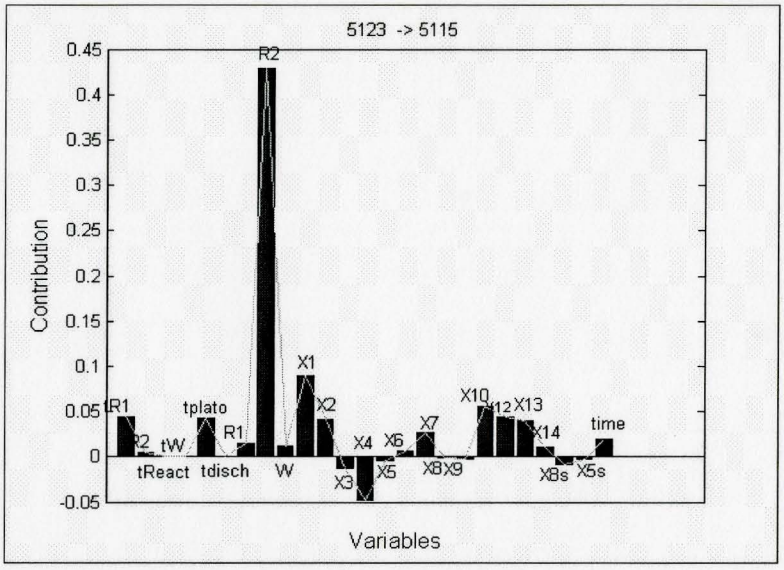

Figure 3.14 Absolute contribution plot for the first PC between batches 5123 and 5115.

A time series plot of the quality data (Figure 3.15) shows that all the "good" batches occurred in the beginning of May. When inquired about this, and the fact that all such batches had considerably higher quantities of reagent 2, the operators of the LX plant confirmed that there had been a problem with the flow meter during this period.

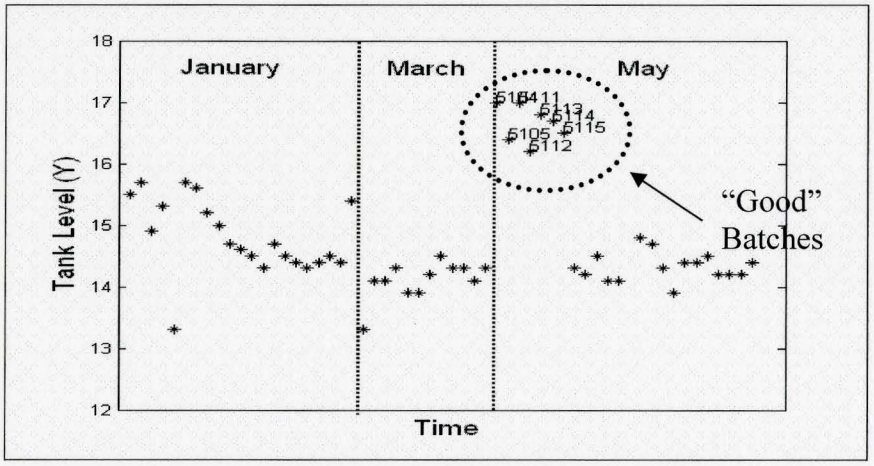

Figure 3.15 Time series plot of the final quality (Y) for the LX data. Production campaigns are discriminated by month.

Retrospectively, the influence of the weight of reagents 1 and 2 added to the reactor on Y values can be seen univariately (Figure 3.16). It is clear, however, that the latent models provided crucial aid at identifying the assignable cause related with the production of the "good" batches.

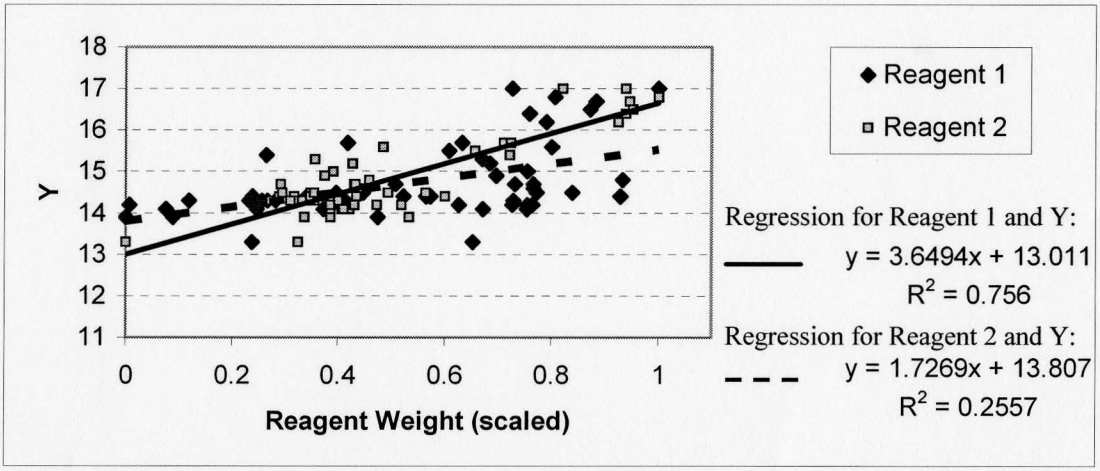

Figure 3.16 Univariate influence of total weight of reagents added and Y. The values of reagent weight has been scaled so as to not disclosure proprietary information.

MB-MPLS model with the exclusion of the cluster of "good" batches

By identifying the special cause that led the batches produced in early May to have higher Y values than all other batches, the standard deviation of Y around it's mean decreased from $\pm$ 0.90 to $\pm$ 0.54 tank level units. In other words, 36% of the variance of Y1, for the given data, is due to a loose control of Reagent 2 addition in batches 5104, 5105, 5111, 5112, 5113, 5114 and 5115.

Through the exclusion of these batches, which do not represent "normal" operating conditions, it is possible to determine which process variables are most influential in explaining the remaining variability in Y.

A 2 PC model (as determined through cross-validation) captured 82% of the variance in Y, 20% of X and 21% of Z. The first latent variable alone is responsible for capturing 73.5% of the variance of Y.

The TT1/TT2 super-score plot for this model indicates that the system is still observable since the batches with highest Y values (mostly produced in early January) have the largest TT1 values, while those with the lowest Y values (mostly produced in March) have the values of this statistic (Figure 3.17).

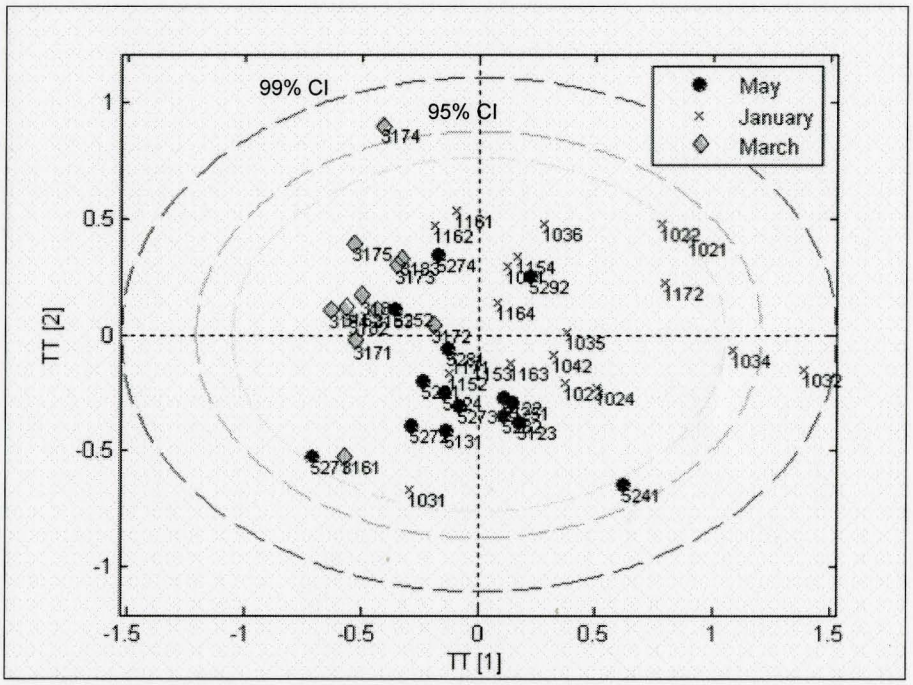

Figure 3.17 Super-score plots TT1/TT2.

Thus, in order to obtain high volumes of the final product (Y), values of TT1 and, consequently, of t1 for the individual blocks ($t1_X$ and $t1_Z$), must be maximized. Analysis of the weights for the first PC corresponding to the X block ($w1_{pt}$) shows that maximization of $t1_X$ occurred when variables X1, X3, X8s and time were kept above average while X5 and X5s where kept bellow average throughout the whole batch time (Figure 3.18). Other variables, such as the bulk (X2, X12 and X13) and brine (X4) temperatures, had time-shifting contributions to keeping $t1_X$ high.

The high relative value of the weights ($w1_{pt}$) for total reactor weight (X1) and it's persistency, makes this variable the main contributor to obtaining high Y values. However, inspection of $w1_{pt}$ is not sufficient to discriminate which of the reagents has the

largest impact on the increase of Y. It is thus necessary to inspect the loadings for the Z block ($w1_{ic}$) for the first PC. Examination of Figure 3.19 shows that high values of Y are obtained when above average amounts of reagent 2 and reagent 1 are added to the reactor. Furthermore, this plot shows that reagent 2 has a bigger impact on Y than reagent 1.

This last observation is in agreement with theoretical knowledge of the process since reagent 2 is known to be the critical, yield determining, raw material in LX production.

The following conclusions were also reached from the inspection of Figures 3.18 and 3.19:

- Even though the cumulative total flow of reagent 2 (X8s) is positively correlated with Y, it's correlation is much smaller than that of the total weight of reagent 2 added (R2). This indicates the need for the re-calibration of the flow meter. It is also worth questioning if, perhaps, a tighter control on reagent addition might be achieved through the use of the reactor scale.
- Brine flow (X5) is shown to be negatively correlated with Y. However, this does not hold true for batches produced within the same campaign (verification was done through the inspection of contribution plots between same-month batches), indicating that such effect is seasonal and correlated with Y due to the fact that the now "good" batches were produced in the colder month of January.

  If the temperature of the in-flowing and out-flowing coolant was measured, than the heat removed from the system could be calculated and no seasonal clustering would be observed.
- Variables X2, X3, X4, X6, X12 and X13 all show positive correlations with Y. This positive correlation is also seen in same-month contribution plots. This indicates that, with relation to quality, these effects are not seasonal. Due to the non-casual nature of empirical models, it is not possible to state if

higher than average temperature values lead to higher Y values or if, due to the fact that more reagents 1 and 2 were added, these temperatures were higher.

This last statement is consistent with theoretical process knowledge since it is known that the LX reaction is exothermic. Thus, higher quantities of the critical reagent leads to higher conversion and, consequently, to higher heat generation.

- Time usage during the batch is positively correlated with Y for the initial stages and negatively correlated for the discharge stage. The fact that time of discharge should be short concurs with theoretical process knowledge that some product decomposition occurs if the material is kept in the reactor too long. However, the stage that is being depicted as having the highest influence on Y is the one in which addition of reagent 2 occurs. This could indicate that a higher time spent on that stage would allow for more material to react or, alternatively, it could just be a consequence of the fact that by adding more reagent 2 more time is spent on such phase.

  Plant experiments in which the reaction phase is prolonged are suggested as a mean to possibly increasing final product volume (Y).

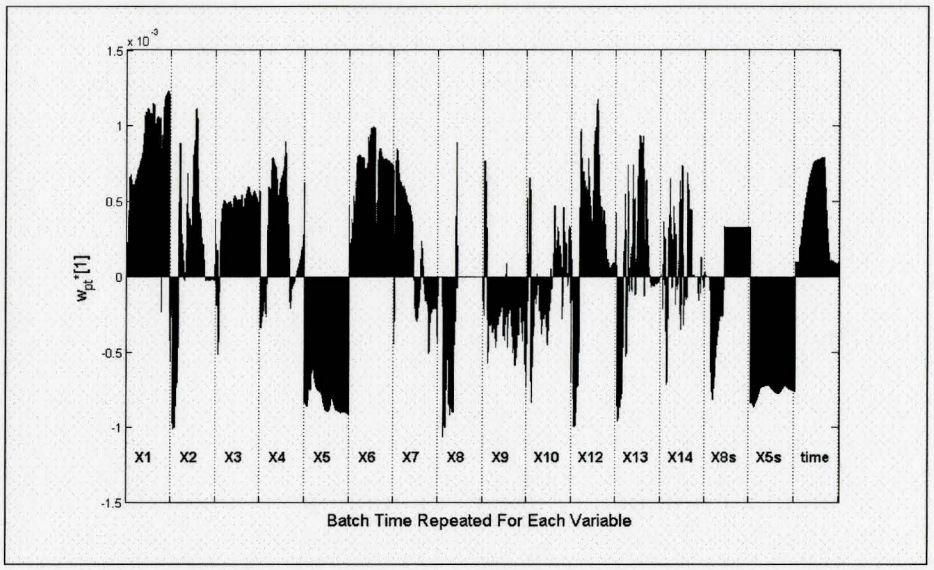

Figure 3.18 Weights for the first component in the X-space for the MB-MPLS.

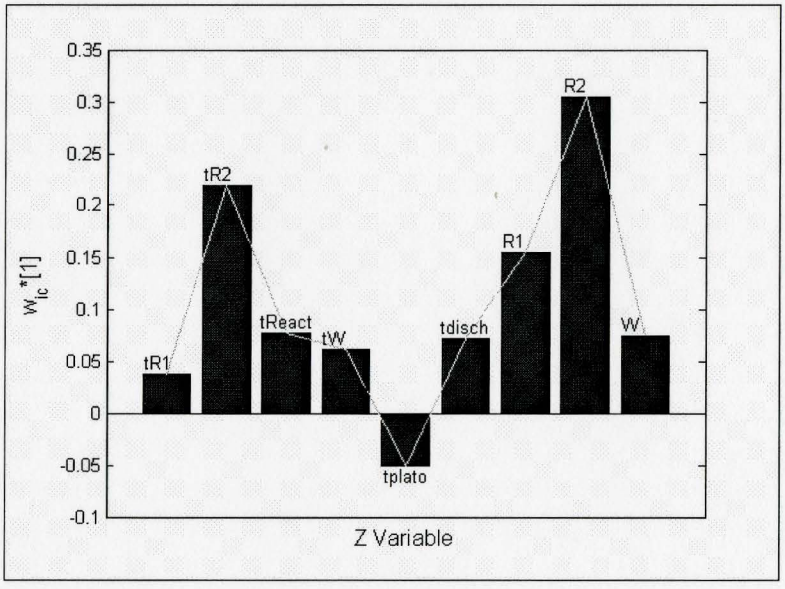

Figure 3.19 Weights for the first component in the Z-space for the MB-MPLS.

From the discussions regarding the potential scenarios for the LX process behavior, it is clear that the amount of reagents 1 and 2 added per batch play a major role in the final volume of product obtained (Y). Due to this fact, their behavior is modeled mostly by the first latent variables of the MB-MPLS model. In order to uncover further

variations in process variables that affect the variability in the Y data it is necessary to inspect the loadings of higher PCs. This, however, can sometimes be difficult to interpret since these effects maybe spread out among various latent variables.

In order to facilitate the analysis for further process variables that impact Y and are uncorrelated with reagents 1 and 2, data laundering techniques will be applied to the LX data set.

3.5.3 Data Laundering

Data laundering or regression out method is a data pre-processing technique that removes "nuisance" or target effects of specific variables from the remaining data set, or from a selected group of variables. Thus, this method allows for the investigation of low frequency signals, uncorrelated with the "nuisance" variation. Examples of such un-wanted variations are: ambient temperature, seasonal and daily effects and reactant concentration streams (Zavitsanou, 2002).

Various methods of data laundering are discussed in the literature (Zavitsanou, 2002). Target rotation (Christie 1995 and 1996) can be used to remove the effects of the target variable through constrained principal component decomposition. The constraint is set on the loading value of the target variable, which is forced to unity. The target rotated model matrix contains all the variations within the data set that are correlated to those of the target variable. The laundered data are contained in the residual matrix (Christie, 1995); the values of the target variable column within this matrix are zero.

Within the least squares solution (Zavitsanou, 2002), initially the laundering coefficients can be estimated by regressing the data matrix (X) on the target variable (x_t) and, in sequence, the laundered matrix is calculated by subtracting the model matrix from the original matrix (residual matrix):

$$\hat{\beta} = (x_t^T x_t)^{-1} x_t^T X \qquad (3.3)$$

$$X_{laundered} = X - x_t \hat{\beta} \qquad (3.4)$$

If more than one target variable exists, Yoon (2001) suggested the use of PLS decomposition to overcome potential problems caused by ill-conditioning of the target matrix:

$$X_{laundered} = X - \hat{X}_{Pedicted\ by\ PLS\ model} \qquad (3.5)$$

Application of data laundering techniques to batch processes must take into account their time-varying nature. To the best of the author's knowledge this specific issue has not been previously discussed in the literature.

The method proposed to launder out a time-varying batch process variable from the remaining data set is to build one regression model for each sample or time point (k) of the batch-wise unfolded data matrix. This concept is illustrated within the least squares framework in Figure 3.20. With relation to this example, xt_k is the target variable and X_k is the process data matrix at time k. The final laundered matrix is calculated by subtracting the model matrix from the original matrix for all time points.

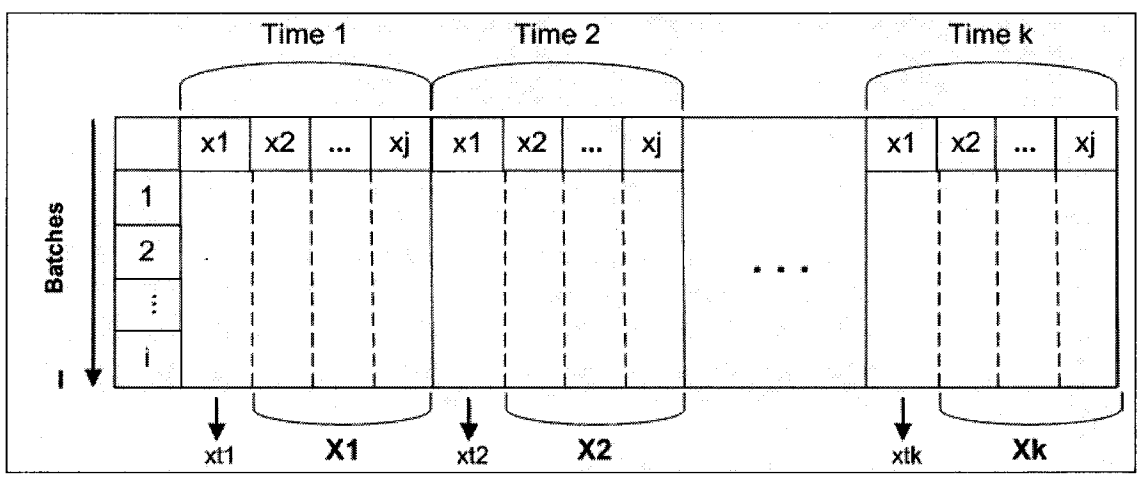

Figure 3.20 Illustration of the method proposed to launder out a time-varying batch process variable from the remaining data set. One regression model is built for each time sample point (k) of the batch-wise unfolded data matrix, following the notation given in equations 3.3 and 3.4.

In the case of the LX process, the "nuisance" or target variables are the total amounts of reagents 1 and 2 added per batch (R1 and R2). If total reactor weight (X1) is set as the target variable and the batch-laundering technique previously suggested is applied, the effect of quench water addition will also be laundered out of the remaining data set and this is not of interest. Thus, the method chosen to selectively remove the effects of R1 and R2, was to apply a PLS decomposition only to the Y data, using the total sum of reagents 1 and 2 per batch as the X data:

$$Y_{laundered} = Y - \hat{Y}_{predicted\ by\ PLS\ model} \qquad (3.6)$$

The PLS model in question was capable of explaining, and thus excluding 46.8% of the variability of Y with one principal component.

Visual inspection of the time series plot of the residuals of the Y data ($Y_{laundered}$) shows that these present a overall decreasing trend. Batches 1024, 1035 and 1034 are considered as being "good" and 3171, 1031, 3161 and 5282 "bad" batches (Figure 3.21).

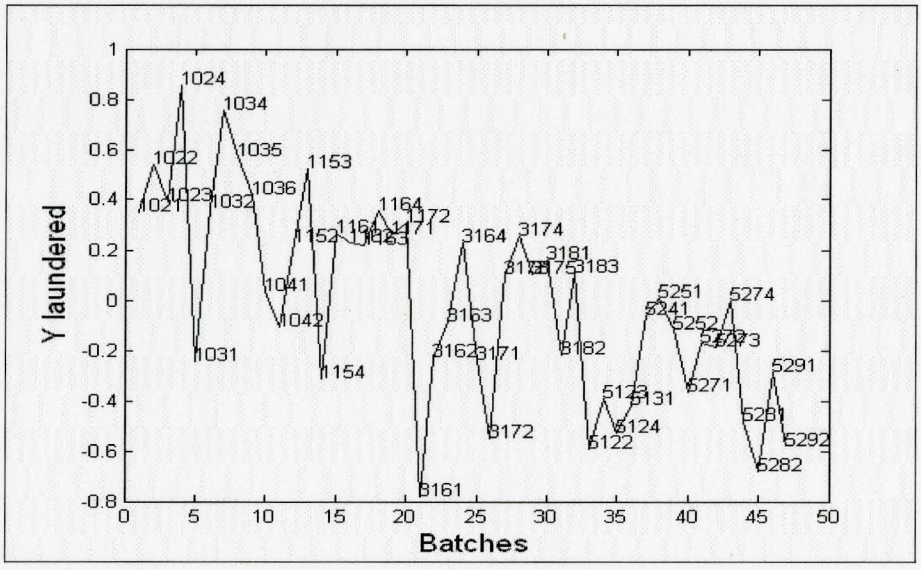

Figure 3.21 Time series plot of the residuals of the Y data ($Y_{laundered}$).

An MPLS model fitted to the LX process data X and the laundered quality data $Y_{laundered}$, was subsequently built using 2 latent variables (indicated by cross-validation).

The Z matrix was kept out of the model since it no longer is needed to discriminate the weight of the different reagents added. The final model was cable of explaining 66% of the variability of $Y_{laundered}$. This means that, with relation to the original data set, the variance of Y explained by the first principal component is 0.066 level units^2 or, equivalently, a standard deviation of $\pm$ 0.25 level units.

Visual inspection of the weights for the total reactor weight (X1) for the first principal component shows that this variable is not correlated with $Y_{laundered}$ for the first two-thirds of the batch (Figure 3.23). This proves that the data laundering technique applied was successful at eliminating the effects of R1 and R2 from Y and, consequentially, from the remaining data set.

However, inspection of the t1/t2 score plot (Figure 3.22) shows that the system is only slightly observable. In general, batches produced in January have higher values of $Y_{laundered}$ and t1 values, however, one is not able to distinguish, between batches 1024 (good) and 1031 (bad) with relation to this last value. Furthermore, clustering among production campaigns is very obvious. This indicates that variations in the quality variable ($Y_{laundered}$) are not significantly explained by variations in the process variables.

Since the variance of $Y_{laundered}$ corresponds to only 8% of the original Y variance it is suspected that information content of the quality data is extremely low and further analysis will not yield conclusive results.

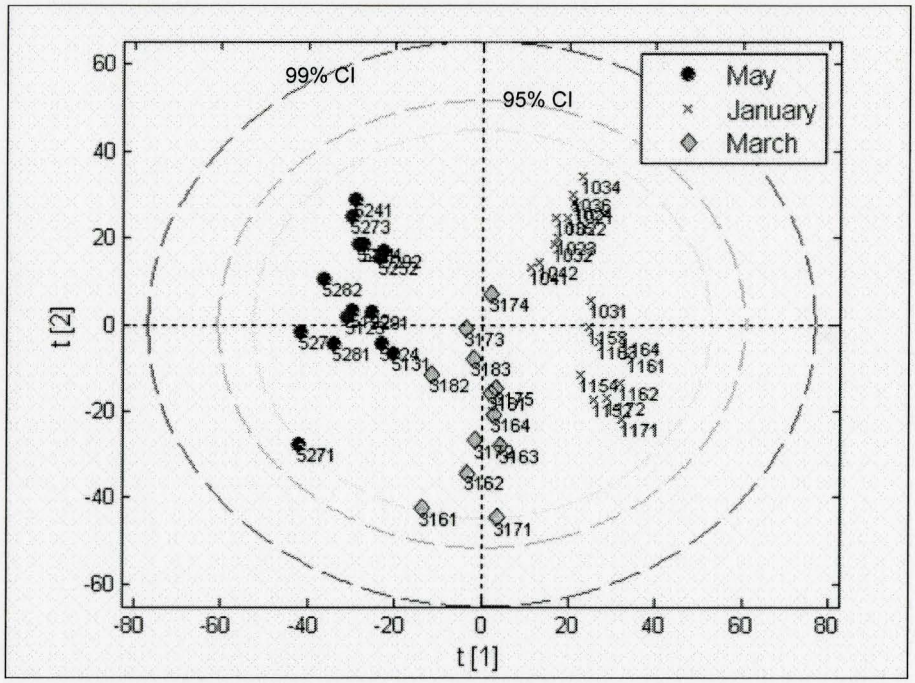

Figure 3.22 t1/t2 score plot for the laundered data set.

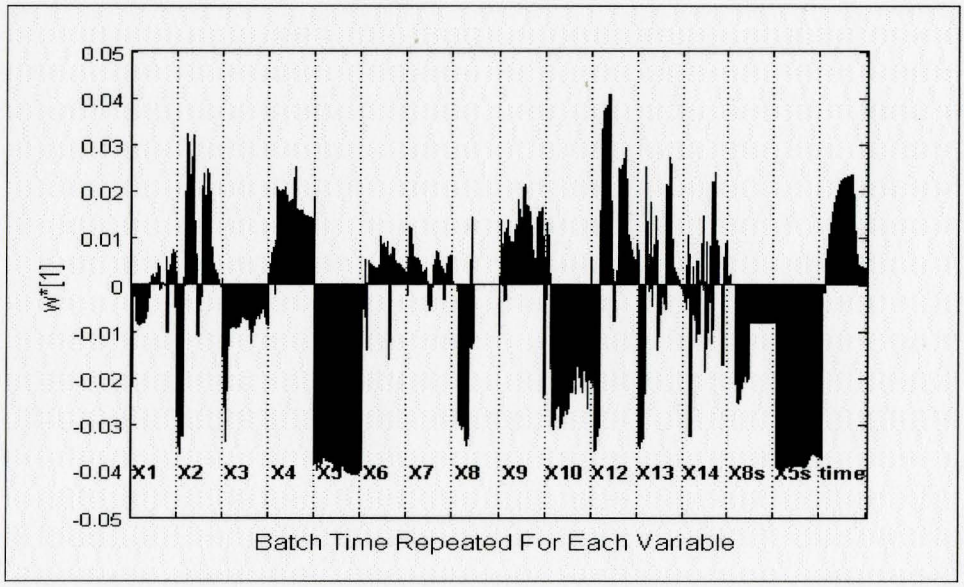

Figure 3.23 Weights for the first PC of the MPLS model.

3.6 Conclusions

Multiway latent variable methods (MPCA and MB-MPLS) were successfully used to troubleshoot the industrial batch process under study. These methods were able to indicate that variations in the total weight of reagent 2 (and to some extent also of reagent 1), added per batch of LX produced, were highly influential in the final product quality (Y). Thus, an increase in the amount of critical component added will lead to an increase in LX liquid volume obtained per batch. Also, reduction of variability in Y can be obtained through a tighter control of reagent addition. Calibration of the flow meter or the use of the reactor scale as an aid in controlling reagent addition is suggested.

With relation to time usage, the latent variable methods indicated that an increase in the reaction time between reagents 1 and 2 (phase 4 shown in Figure 3.4) and a decrease in the time of discharge may lead to an increase in Y values. Plant tests are needed to further determine the causal nature of this observation.

Additionally, the off-line MPCA analysis performed was capable of identifying a seasonal behavior in some of the LX process variables caused by changes in ambient temperatures.

Data laundering techniques specific to batch processes were also suggested, so that further variations in the data set could be explored. However, their application to the current process led to inconclusive results since the information content in the data was low.

Finally, an observation is made to the fact that the final quality data metric chosen for the LX unit (total product volume obtained per batch) is, most likely, not capable of sufficiently describing the true final product quality for this process. For an effective PLS analysis, the Y matrix should span a wide range of product properties: physical, chemical and relative to final application (Nomikos, 1995). In the case of the process under study measurements of the percentage of active material present in the final product, for example, would probably allow for greater process understanding, optimization, and possibly, mid-course corrections.

Chapter 4

On-Line Monitoring of a Multi-Grade Batch Annealing Process using MPCA

The purpose of the current Chapter is to present a novel, multi-grade, industrial application of the on-line multivariate monitoring methodology introduced by Nomikos and MacGregor (1994) for single-grade batch processes. More specifically, in the current work, this new methodology was successfully used to build a single, all-encompassing, on-line monitoring scheme for the heating phase of a multi-grade batch annealing process.

4.1 Batch Annealing Process Description

Annealing is an important heat-treatment process used to soften, relieve stress and increase ductility of cold-rolled steel (Moon and Hrymak, 1999). Physically this is achieved through mechanisms of recovery, recrystallization, and grain growth of deformed metal microstructures (Sahay and Kumat, 2002).

Annealing can be either a continuous or a batch type operation. Due to their versatility, economics and ease of operation, the majority of the existing annealing facilities, including that under study, rely on batch type processes (Moon and Hrymak, 1999).

Batch annealing operations take place on a fixed base whereupon cylindrical steel coils, separated by convector plates, are stacked. This system is then enclosed by a

protective cover, followed by a furnace. An inert gas is circulated within the protective cover through the aid of a base fan. Depending on the type of fan installed, two different speed settings can be fixed throughout a batch run. Similarly, there are two different furnace types currently in use at the steel mill in question.

Three temperature sensors (T1, T2 and T3) are located at different points of the annealing reactor and a fourth temperature measurement (T4) is inferred from the data collected by these sensors. A schematic of the process described is shown in Figure 4.1 and typical temperature profiles for a single processing cycle are given in Figure 4.2.

Each annealing cycle is composed of a heating and a cooling phase. The beginning of the heating phase is marked by the ignition of the furnace burners followed by a fast rate of increase in T1. When this variable reaches it's set point value, soaking begins. The simultaneous end of the soaking step and the beginning of the cooling phase occurs when T4 reaches set point value. At the beginning of the cooling phase, the furnace is removed, with the aid of a crane, and replaced with a cooler.

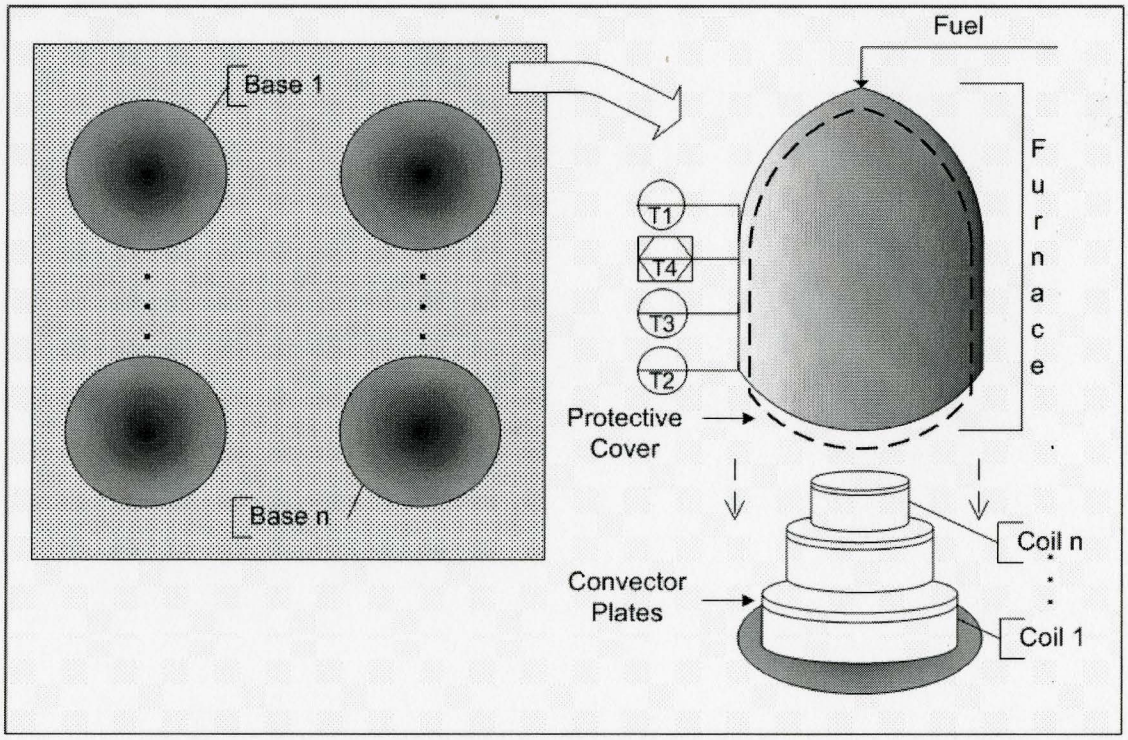

Figure 4.1 Schematic of the batch annealing process.

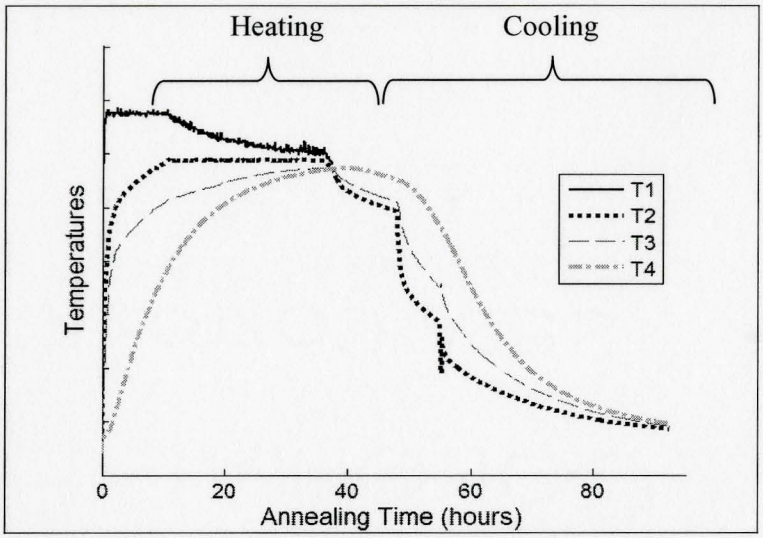

Figure 4.2 Typical batch annealing temperature profiles for a single processing cycle. T1 values are not registered during the cooling phase.

The existence of multiple bases, protective covers and furnaces, allow for different stacks of steel coils to be annealed independently. Each stack can contain from 1 to 4 coils of the same steel grade. A total of 11 steel grades are processed at the facility under study.

A complete annealing cycle takes anywhere from 3 to 7 days (Moon and Hrymak, 1999). Soaking times and temperature set point values depend on steel grade and are generally pre-determined through plant trials and empirical methods (Sahay and Kumat, 2002). During production, overall processing times and temperature trajectories are also subjected to batch-to-batch variations mainly depending on the number and geometry of the coils being annealed and, to varying degrees, on process equipment specifications (furnace type, base fan speed and base location). This is further discussed in Chapter 5 for one particular grade type.

4.2 Project Incentives and Objectives

High temperatures and long heating phase cycle times (30 to 60 hours) make batch annealing operations energy intensive (Sahay and Kumat, 2002). This issue, which has a large impact on processing cost, associated with high final product quality demands and a growing number of new steel grades, have led to a revival of interest in batch annealing studies.

Perin *et al.* (1988) discuss the factors that have a direct effect on the performance of batch annealing equipment (mainly high convection heat transfer coefficients within the protective cover). Buckley *et al.* (1999) used sophisticated computational fluid dynamics (CFD) techniques to simulate the heat transfer phenomenon occurring in a high temperature coil annealing furnace (HTCA). Both of these technologies can be applied to the design of annealing furnaces to greatly increase productivity of conventional equipment and final product quality. However, save re-design, improvement of already installed processes is not addresses by these studies.

Moon and Hrymak (1999) addressed short-term scheduling issues for batch annealing processes with the objective of maximizing material throughput. Within this work a novel mixed-integer linear programming (MILP) model was successfully used to determine the optimal equipment movement plan capable of satisfying high utilization of shared equipment (i.e. cranes, furnaces and coolers). The general mathematical model developed for this optimization study is deterministic in nature and assumes, among other things, that no failures in utilities occur during an annealing cycle. The authors recognize that this assumption is not reasonable in real batch processes and that a further gain in process efficiency can be achieved through real-time fault identification.

Shay and Kumar (2002) developed an integrated batch annealing furnace simulator (BAFSIM) capable of predicting temperature evolutions, microstructural and mechanical steel properties. The main purpose of this simulator is to optimize annealing process cycles in order to achieve an optimal balance between high productivity and high final product quality with a reduced number of plant trials. Additional potential uses of

this model include: online control, soft sensor, product development, trouble-shooting, statistical process control and information management.

Although the models of BAFSIM are based on fundamental relations, tuning parameters are determined through the use of plant data. However, no description was given by the researchers regarding the number of batches necessary to accurately determine these tuning parameters and which factors, such as staking configuration and steel grade, impact them.

The objective of the current project is to build a single, all-encompassing, data-based, on-line monitoring scheme for the heating phase a multi-grade batch annealing process. Such scheme should, not only be capable of real-time fault detection for the various steel grades processed at this site, but also be able to handle the various coil staking configurations, furnace types, base locations and fan speed settings used.

Due to it's ease of use and capability of handling large sets of noisy, correlated, batch process data, multiway principal component analysis (MPCA) is the empirical method of choice in building this monitoring model.

Implementation of an efficient on-line monitoring scheme allows for early fault detection and, consequently, for faster corrective actions by the operators. Such actions can greatly improve efficiency of material throughput and process safety. Additionally, batch-to-batch final quality variability is reduced through the elimination of abnormal processing conditions.

The advantage of having a single, all-encompassing, model (as opposed to one model for each steel grade) is the reduction of implementation and upkeep efforts. Champagne and Ivanov (2002) listed the necessity of building one model for each product grade produced in a given process as a major limitation of the use of multivariate analysis tools. These authors developed a multi-grade modeling technique which uses a PLS discriminant analysis (PLS-DA) model to extract and eliminate inter-grade variability from a continuous paperboard production dataset. The PLS-DA residuals are then used to capture intra-grade variability with a second PCA or PLS model. In

summary, this methodology is able to group families of similar, continuously produced, paperboard grades with two models.

To the best of the author's knowledge, the current work is novel in the sense that it presents a multi-grade modeling technique, for monitoring batch processes.

4.3 Description of the Data Set

The data set provided by the annealing facility to build the reference PCA model included a total of 83 batch runs. Each batch run contains 4 process variables (T1, T2, T3 and T4) which are sampled at 5-minute intervals, during both the heating and cooling phases of the process. A total of four different steel grades (G1, G2, G3 and G4) are represented within this data set, corresponding to 72% of all steel annealed in this facility. For each steel grade, staking arrangements of both 2 and 3 coils were included. Due to lack of data, arrangements of 1 and 4 coils were not included in the model set but were subsequently tested.

In order to further test the monitoring scheme, data for 33 additional batch runs was provided. This new data set contained both batches under normal process operation (8 observations) and batches in which a fault had occurred (25 observations). The data pertaining to the "out-of-control" batches was carefully selected to contain examples of the most common faults (section 4.5.2) for all the different steel grades encompassed by the monitoring scheme. The new "in-control" data was purposefully collected months after the data used to build the monitoring scheme with the intent of picking up any slow drifts in the process that may have occurred throughout this time.

As previously discussed, annealing temperature trajectories are subjected to batch-to-batch variations depending on steel grade, number and geometry of the coils being annealed and on process equipment specifications (furnace type, base fan speed and base location). It is important that an approximately equal number of batches with different combinations of such parameters be included in the reference data set. This is

done to avoid that the monitoring scheme be biased towards specific trajectory characteristics.

Table 4.1 shows all the data sets used throughout this work as well as their distribution with relation to steel grade and staking configurations. Distribution with relation to all other parameters previously listed was considered satisfactory.

Table 4.1 Description of the data sets provided.

Data set purpose	Total Number of Batches	Number of batches with 1,2,3 or 4 coils per steel grade			
		G1	G2	G3	G4
Reference model building "in-control"	83	0,5,15,0	0,10,7,0	0,9,16,2	0,12,7,0
Model testing for 1 and 4 coils "in-control"	6	1,0,0,2	0,0,0,1	0,0,0,2	0,0,0,0
Model testing for process drifts "in-control"	8	0,0,2,0	0,2,0,1	0,1,1,0	0,2,0,0
Model testing for faults "out of control"	25	0,3,3,0	0,1,5,0	0,1,5,1	0,3,3,0

4.4 Data Pre-treatment

The steps taken in order to treat the batch annealing data set, prior to building the final PCA monitoring model, are described in the sub-sections that follow. The results of all individual steps are shown in Figure 4.4 for variables T1 and T4.

4.4.1 Data Visualization and Trimming

The first step taken towards data analysis was the graphical representation of all batches with the purpose of process behavior visualization (Figure 4.4, top, shows the raw data for T1 and T4). This preliminary analysis revealed that four batches presented outlying starting values of T1. These abnormal data points were substituted for missing data after plant personnel confirmed that they were in fact a result of sensor failure at such points.

In sequence, the heating phase of each batch run was identified based on inferred temperature measurements (T4). All batch cycles start at the same T4 value. The end of the heating phase occurs when the set point for T4 is reached. Since the objective of the current project is to monitor only the heating phase, all data not pertaining to this interval was trimmed and discarded.

4.4.2 Data Alignment

Due to variations in the time necessary for T4 to reach set point value, annealing heating cycles have different durations from batch to batch. One method of guaranteeing that all batch runs have the same number of samples and matching trajectories is to align the data by resampling it based on intervals of one or more indicator variables.

For on-line monitoring purposes, it is necessary to choose a monotonically increasing indicator variable with known initial and final values. In the current data set, the only variable that satisfies this condition is T4. However, this variable is not monotonically increasing in the following situations: (i) the initial and final sections of most batch runs (ii) cases where a specific type of fault occurs causing this temperature to drop. In order to overcome these problems, the following procedures were adopted during the resampling portion of this work:

1) The first 15 samples are kept at 5 minute intervals to guarantee that important variations in initial process measurements are captured.
2) The value of T4 at the 16^{th} sample of each batch is set as being equal to 0% of batch completion. The final set point value of T4, which depends on the grade of steel being annealed and is known before the start of the batch, is set as being equal to 100% of batch completion.

   Samples were taken at every 0.25% increase of T4 between 0% and 99% of batch completion and at every 0.05% increase of such variable for the

remaining of the batch run. This multi-rate sampling approach is done to compensate for the slower raise of T4 at the end of the batch.

For the sample rates chosen, all "in-control" batches had samples taken at a maximum of 5 minute intervals for the first half of the batch, 15 minute intervals from half to three-thirds of batch completion and 30 minute intervals in the remaining section. These sampling rates were considered acceptable by plant personnel. It is important to remember that, for the current process, data samples come from the process at 5 minute intervals and thus, resampling at smaller intervals will only lead to data interpolation and not acquisition of new information.

3) In order for new samples to be taken in cases where T4 values drop, a condition was added to the aligning program guaranteeing that, if a sample had not been collected by the monitoring scheme at time intervals superior to the maximum values exposed in the paragraph above, a sample is taken, independent of T4 values.

The results of this aligning technique can be seen in Figure 4.4 (T1 Aligned and T4 Aligned).

4.4.3 Unfolding, Grade Specific Mean Centering and Scaling

According to the theoretical discussion presented in Chapter 2, the next logical step to multiway, multivatiate latent variable data analysis is to mean center and scale to unit variance the full, batch-wise unfolded, process data matrix. In other words, the statistical parameters (row-wise means and standard deviations) necessary for this mathematical operation are calculated based on the reference data set which, in this case, is composed of 4 different steel grades.

A MPCA model, composed of two latent variables, was fitted using data pre-processed using this standard approach. Visual inspection of the resulting t1/t2 score plot

(Figure 4.3, left) shows a clear clustering of batches in which the same steel grade was annealed.

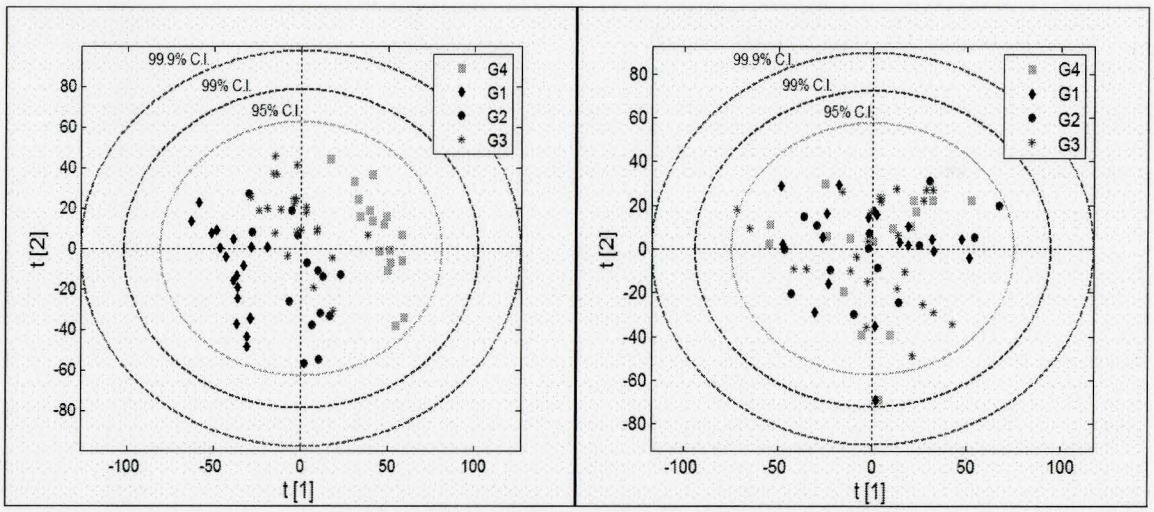

Figure 4.3 t1/t2 score plot for conventionally mean centered (left) and grade specific mean centered (right) batch data.

This behavior is expected since temperature set points and processing times are grade dependent. Thus, batches where the same steel grade is annealed have similar behavior and are thus projected within the same cluster in the score plot. A direct consequence of this segregation is a widening of the statistical control limits calculated based on the reference data set. The monitoring scheme then becomes less sensitive in terms of fault detection.

Kosanovich *et al.* (1999) showed how independent scaling of data sets collected from different reactors can be used to remove the segregation between them. A similar concept can be applied to the multi-grade annealing process in the following manner:

1) Initially the reference data set matrix (X) is divided into sub matrices (X_{Gi}) containing only process data for those batches in which the same steel grade was annealed.

2) The mean (row-wise) of each unfolded sub matrix (X_G) is then calculated and used to perform grade specific mean centering of batches that have similar steel grades. On-line mean centering of new batches is possible using this approach since the steel grades to be annealed are known ahead of time.
3) Posterior of grade-specific mean centering, the sub matrices (X_G) are once again united to form the full reference set (X), which is scaled to unit variance.

The results of this pre-processing technique can be seen in Figure 4.4 (T1 and T4 mean centered). By comparing the data that has been simply aligned to that that has also been mean centered by steel grade; one observes that no segregation of the temperature trajectories occurs when this later technique is employed. Additionally, Figure 4.3 (right) shows that the grade specific mean centering technique applied was also successful at eliminating data clusters in the t1/t2 score plot.

Additionally, grade specific scaling could also have been applied to this data set to further decrease grade-to-grade clustering in the latent variable space. However, all grades presented approximately the same variance for all variables throughout the whole batch run and thus this additional pre-processing step was not considered necessary.

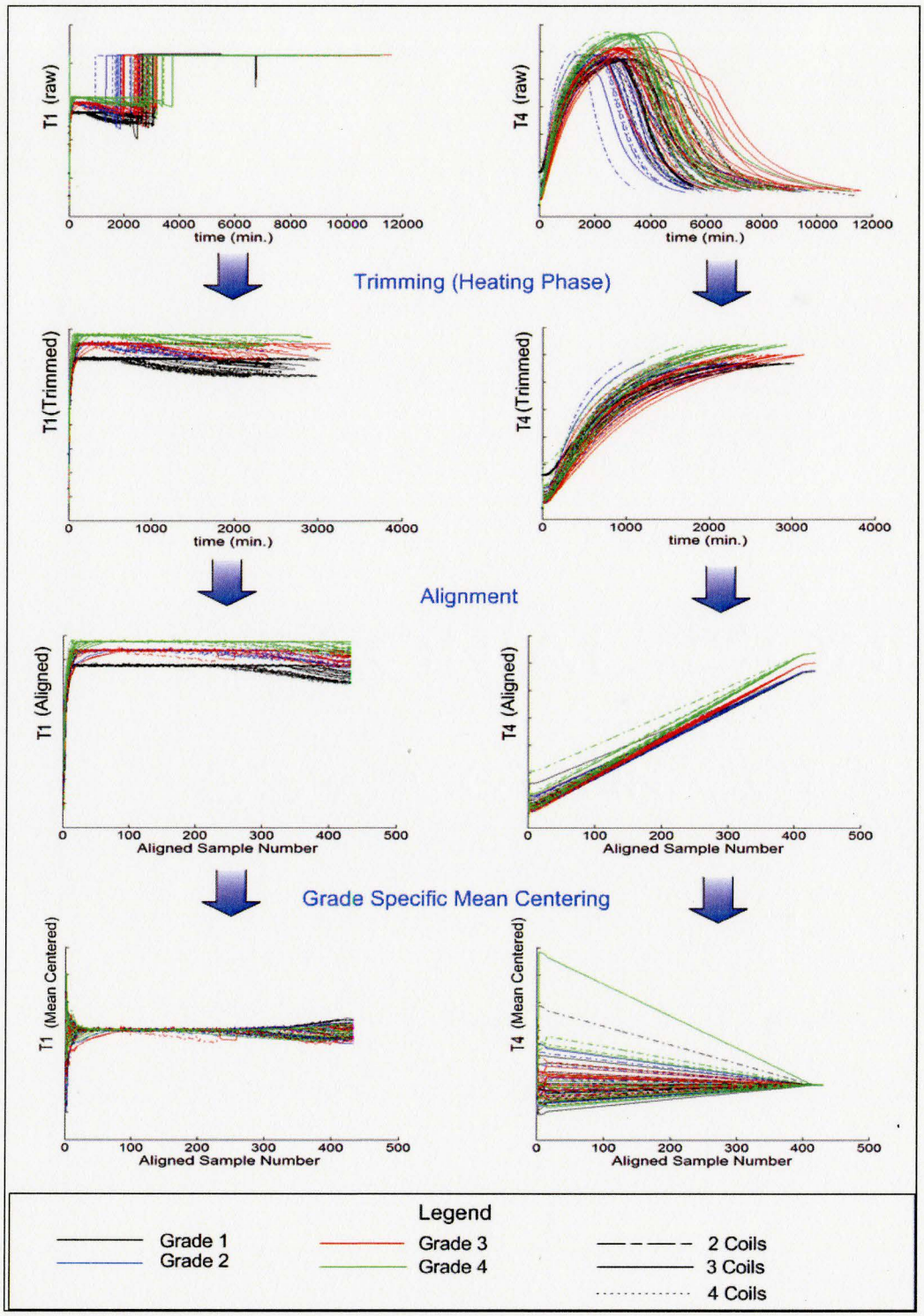

Figure 4.4 Data pre-processing techniques variables T1 and T4 of the reference data set provided.

4.4.4 Data Augmentation

With the purpose of enhancing overall fault detection and identification, cumulative time per batch (Time) was included as an extra variable in the X matrix (Westerhuis *et al.*,1999, Garcia *et al.*, 2003 and Kourti, 2003). In order to improve the detection of faults caused specifically by irregular fuel feed to the furnace, a new variable (T1var) was also incorporated to the original data set. This type of fault is characterized by an increase in the variation of T1 around it's mean value and normally indicates the existence of a sticky valve (see section 4.5.2 for further explanations and a visual representation of T1 and T1var behaviors, both typically and during fault occurrence). T1var reflects the cumulative variance of T1 at every sample point (k) and is calculated by:

$$T1\,\mathrm{var}_k = \frac{(T1_k - \overline{T1}_{(p:k)})^2}{k-p} + \frac{(T1_{k-1} - \overline{T1}_{(p:k-1)})^2}{k-1-p} + ... \qquad (4.1)$$

Where: k = current sample point;

p = initial sample point;

$T1_{(p:k)}$ = T1 average calculated using all points between p and k;

It should be noted that the use of equation 4.1 causes samples k to have a lower weight with relation to previous ones. Alternatively, a variable reflecting the variance at every time point could have been used.

Empirical studies indicated that best fault detection results were achieved when the values of T1var are set to zero during the samples in which T1 has not reached set point value.

Experimentally, models containing two new variables expressing the distance between T1-T2 and T2-T3, at every sample point, were also fitted. However, fault detection capabilities of these models were not superior to that of models without these two variables.

Thus, the final X matrix contains a total of six variables: Time, T1, T2, T3, T4 and T1var.

4.5 Reference MPCA Model

This section describes the steps taken in building the reference MPCA model to which all new batches under production will be compared against, in real-time, and classified as normal or abnormal. Theoretical concepts regarding reference model building are further described in Chapter 2.

4.5.1 Selection of "In-control" Batches

Reference statistical models are built using a historical set of "good" or "in-control" batches. All 83 batches included in the original data set were believed, by plant personnel, to be representative of the batch annealing process while under normal operation. This assumption was made due to the fact that the original monitoring scheme did not alarm during their production.

In order to determine if these batches were truly "in-control", the following iterative process was used: i) a MPCA model, composed of 2 PCs, is built - initially using all batches and subsequently only those remaining after step iii; ii) overall Hotelling's T^2 and SPE statistics were calculated for each batch (these plots provide a diagnostics to test if any unusual batches have been included in the reference data set and if a model that is representative of normal operation has been built [Nomikos and MacGregor, 1995]); iii) batches in which these statistics greatly exceeded their respective 99.9% confidence intervals (C.I.) were excluded; iv) the previous steps were repeated until all outliers were removed from the data set.

At the end of the procedure described above, a total of 7 batches were excluded from the original data set. After visual inspection of the data from these batches, plant personnel verified that 4 of them presented faults that had not been diagnosed by the

original monitoring scheme. The remaining batches were considered as being "fast", due to the fact that all temperatures rose more rapidly than normal. Additionally, some of these batches presented shifts in furnace fuel operation. This, in itself, is not a fault; however, it is not depicted in other batches and thus leads to high overall SPE and Hotelling's T^2 values. Thus, due to their high leverage, all 3 batches were eliminated.

Of the remaining 76 "in-control" batches, 2 were excluded for model performance testing. The final reference process data matrix is thus composed of 74 batches.

4.5.2 Selection of the Number of Principal Components

The optimal number of principal components (PCs) needed to parsimoniously describe the main variations in a given data set, can be calculated using different criteria (see section 2.1). Cross-validation is currently the mostly widely used method for this purpose. The results obtained from the use of various statistical tests in the cross-validation procedure of the pre-processed batch annealing data, along with the software used, are listed in Table 4.2.

Table 4.2 Description of the statistical tests used for cross-validation.

Selection Criteria	Software used	Number of Principal Components selected to build the MPCA model
R	Simca P+ version 10	> 20
R	Simca P+ version 11	10
R	Batch SPC version 2.0	3
W	Batch SPC version 2.0	10
Minimum Press	Batch SPC version 2.0	> 20

It is worth mentioning that, whenever Simca P+ was used during this exercise, the data set was previously unfolded in a batch-wise manner and treated as a normal project. This software has a different procedure for determining the number of PCs that best

describes variable-wise unfolded batch data. If these rules were applied to the current data set, 3 PCs would be selected.

Inspection of Table 4.2 shows that there is a considerable difference in the results obtained for number of PCs selected as optimal for building the MPCA model. These results depend, not only on the statistical criteria that is applied, but also on which version of which software is used. This is due to the fact that additional rules, which vary from software-to-software (Eriksson *et al.*, 1999) and even from version-to-version, are added to cross-validation procedures. These rules determine, among other things, the size of the data set that is kept out from the model at each step of calculations and the significance of the statistics on which the stop point is based.

The main conclusion resulting from this exercise is that statistical tests used for cross-validation procedures are not sound. These criteria only supply a guideline for the selection of the necessary number of latent variables that should be used to build a reference model. In order to best estimate this parameter, it is necessary to consider the purpose of the model and the overall picture that different selection criteria give (Nomikos and MacGregor, 1995a and Eastment and Krzanowski, 1982).

The MPCA model in question will be used for on-line monitoring and, as such, it's main purpose is to be able to distinguish, in real-time, when an abnormal process condition is occurring and when it is not. Charts based on Hotelling's T^2 (using A principal components) and on SPE statistics provide a very effective set of multivariate monitoring techniques (Yoon and MacGregor, 2000). Thus, these chats will be used to detect faults that may occur during an annealing cycle. When a fault occurs, the instantaneous values of the Hotelling's T^2 indicate the occurrence of larger than normal variations in measurements that are consistent with the model, while the instantaneous SPE statistics accounts for disturbances which are not represented in the reference data set and break normal process correlations (Qin, 2003).

Figure 4.5 shows how the multivariate monitoring charts depict the occurrence of a fault during the production of a single batch. In this case both charts were able to detect the fault at around the 25^{th} sample point. This dual detection can occur when the

abnormal behavior is captured to a small extent in the reference database but occurs with large intensity during a fault.

Visual inspection of Figure 4.5 also shows that the size of the C.I. used can influence fault detection; delaying it when a wider C.I. is used. Thus, this parameter must also be taken into consideration during PC number selection.

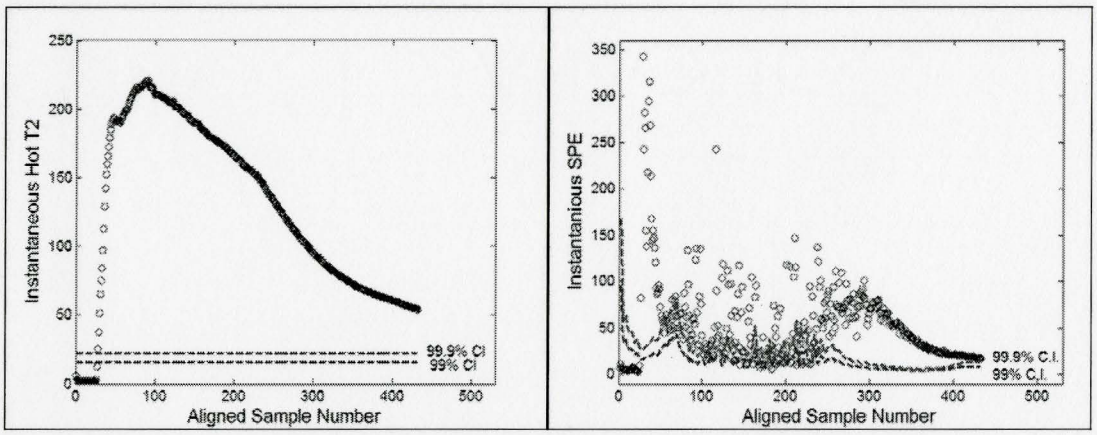

Figure 4.5 Hotelling's T2 (left) and SPE (right) instantaneous monitoring charts for a batch in which a fault has occurred.

The people who will be in closest contact with this model are the plant operators and thus, it is of crucial importance that they be satisfied with it. According to annealing plant personnel, operators rapidly lose confidence on a monitoring scheme if it alarms when no fault is present. Inevitably they start to ignore all alarms coming from this scheme, rendering it useless.

With these issues in mind, metrics relative to delayed or missed fault detection (Type II error) and false alarms (Type I error) are chosen as criteria for latent variable number selection. The objective is to fit an MPCA model such that a reliable monitoring scheme is obtained (both types of errors are within acceptable limits).

In order to indirectly determine the occurrence of Type II errors, a model with a randomly selected, high number, of principal components (11 PCs) was built using the annealing reference data set. In sequence, the number of samples it took for this model to

detect faults occurring in a few "out-of-control" batches was verified (using both Hotelling's T^2 and SPE monitoring charts). The number of latent variables used to build this model was then sequentially reduced and the moment of fault detection was again determined for each new model. The delays of the alarm between the new models and the one built with 11 PCs was determined and are represented in Figure 4.6 (left) for both 99% and 99.9% C.I.

For batch processes, Hotelling's T^2 and SPE values at successive times points are not independent and thus, Type I errors are not equal to the α values resulting from control limit tests applied to these statistics. Nomikos (1995a) suggests a procedure for determining Type I error for the control limits over an entire batch run: i) each batch in the reference data set is passed through the monitoring procedure; ii) the number of statistics (Hotelling's T^2 and SPE values) falling the outside the control limits is determined; iii) the sum of these "outlying" points is divided by the total number of observations (I x K). The result of applying this procedure to the annealing data is shown in Figure 4.6 (right).

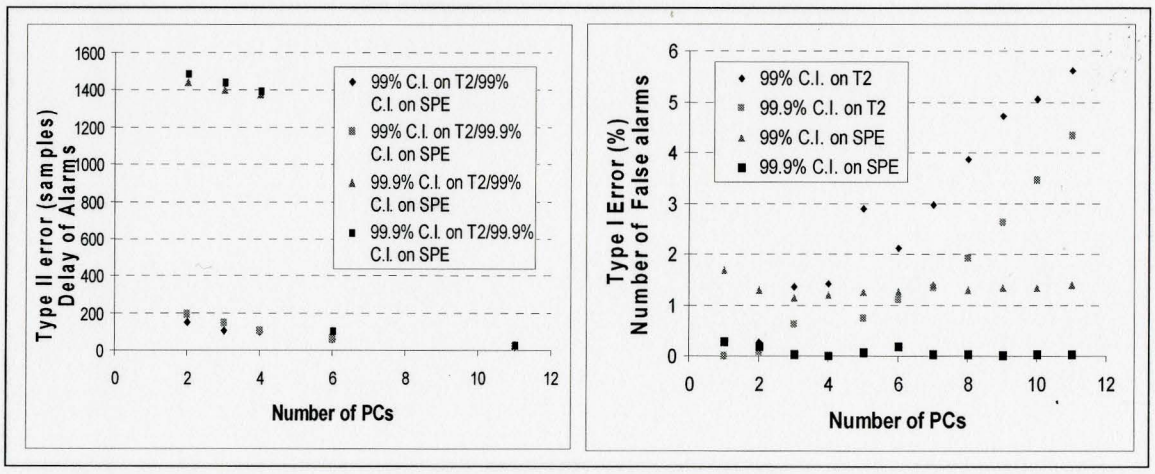

Figure 4.6 Distribution of Type I and Type II errors with an increasing number of principal components and different C.I. values.

Inspection of Figure 4.6 lead to the following observations:

1) Type II errors decrease with an increasing number of PCs, while Type I errors generated by the Hotelling's T^2 chart increases with this parameter. This is due to the fact that the greater the number of latent variables used to build a model, the more this model is capable of capturing behaviors inherent to specific section of specific batches (which, for all practical purposes may be characterized as noise). Thus, these models are more sensitive to deviations from these behaviors, both when they are in fact descriptive of a fault and when they are not.
2) It is interesting to note that the estimated Type I errors for SPE are close to the instantaneous α values for all combinations of PCs and C.I. parameters (α = 1% for a 99% C.I. and α = 0.01% for a 99.9% C.I.). However, with relation to the Hotelling's T^2 statistics, Type I error values become larger and larger then α values for increasing PC numbers.
3) Considering that the annealing process does not present faults that have a diret impact on process safety and that operator confidence on the monitoring system is of extreme importance, plant personnel stated that Type I errors above approximately 1% are not desired.
4) Models built using 4 or less PCs with 99.9% C.I. on Hotelling's T^2 led to missed fault detection of one or more batches. Thus, this combination of parameters cannot be used in the final model.

One way of verifying the statement made that principal components of higher order capture behaviors inherent to specific batches is to analyze the loadings of the each latent variable. Visual inspection of Figure 4.7 shows how the loadings for the first and the fourth principal components are, for most variables, constantly positive or negative throughout the entire batch run. Alternatively, Figure 4.8 (left) shows that the loadings of the seventh principal component have alternating signs throughout the batch run. This is

due to the fact that the 7th latent variable is mainly capturing the behavior of batch 183171; which has the highest leverage with relation to this model. Examination of the temperature trajectories for this batch (Figure 4.8, right) shows that, for the end of the batch, T1, T2 and T3 were low, as were the loadings for these variables.

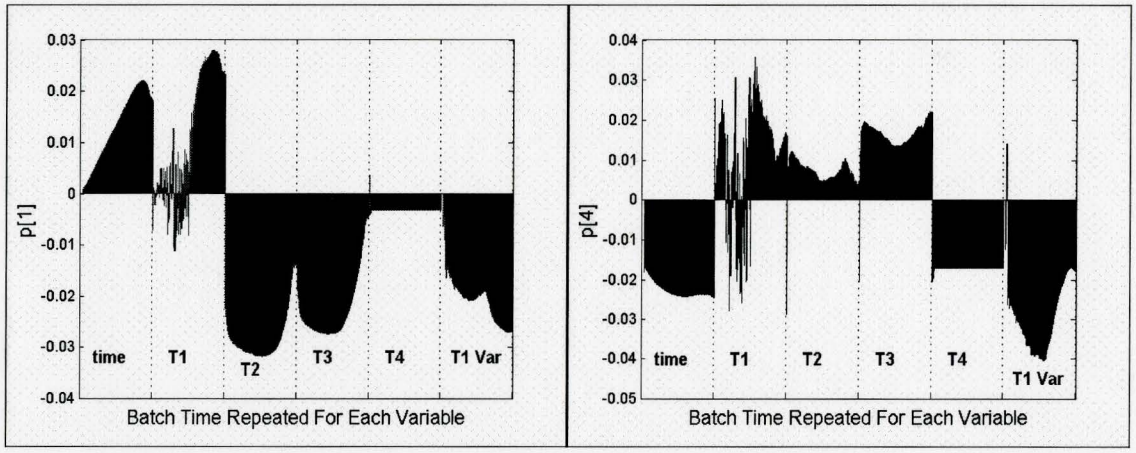

Figure 4.7 Loading plots for all variables at all times for the first (left) and fourth (right) principal components.

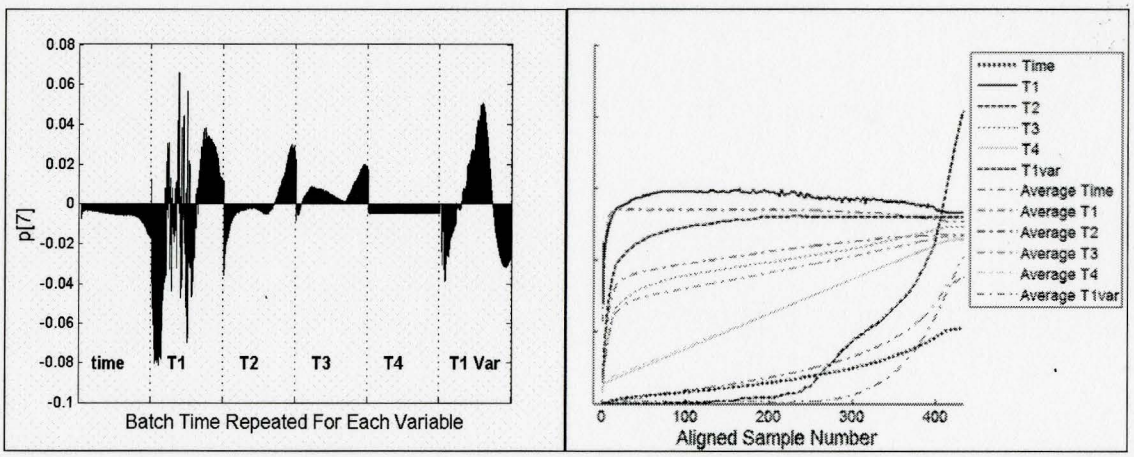

Figure 4.8 Loading plot for all variables at all times for the seventh principal component (left) and process variable trajectories for the batch with the highest leverage in the seventh PC (right).

Figure 4.9 shows R^2 and Q^2 values for models built with a successively higher number of principal components. Close inspection of this plot indicates that, from 4 PCs on, R^2 values do not significantly increase (5% or more) through the addition of an extra latent variable; for Q^2 this occurs from 5 PCs on.

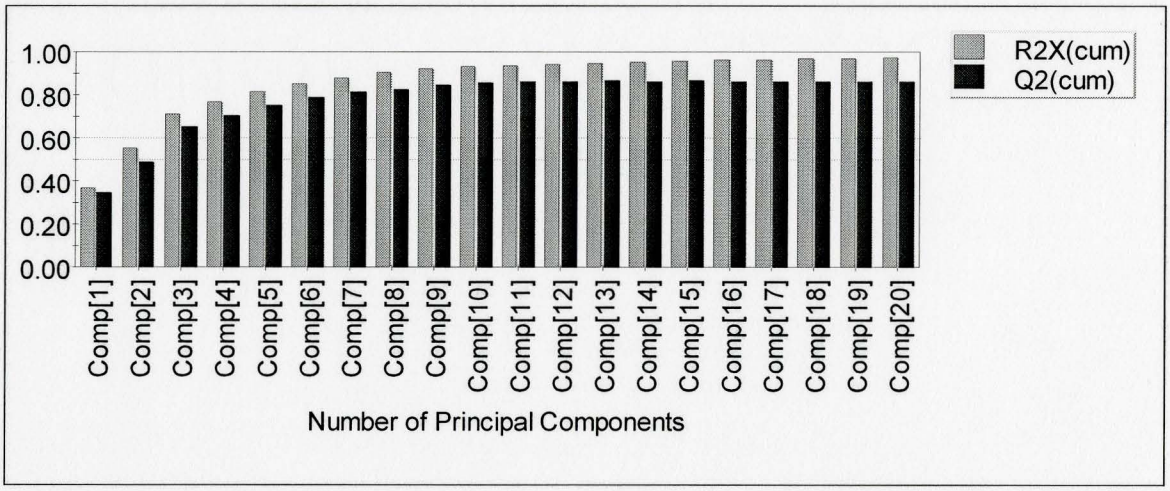

Figure 4.9 R^2 and Q^2 values for models built with a successively higher number of principal components.

Taking all listed observations into account, 4 principal components were selected to build the reference model. Furthermore, a 99% C.I. on Hotelling's T^2 and a 99.9% C.I. on SPE were used.

The final MPCA model, built using 4 latent variables, is capable of capturing 76.4% of the total variability in the data. Overall Hotelling's T^2 and SPE plots (Figure 4.10) show that the outlier detection method applied in the previous section was successful at eliminating all abnormal batches.

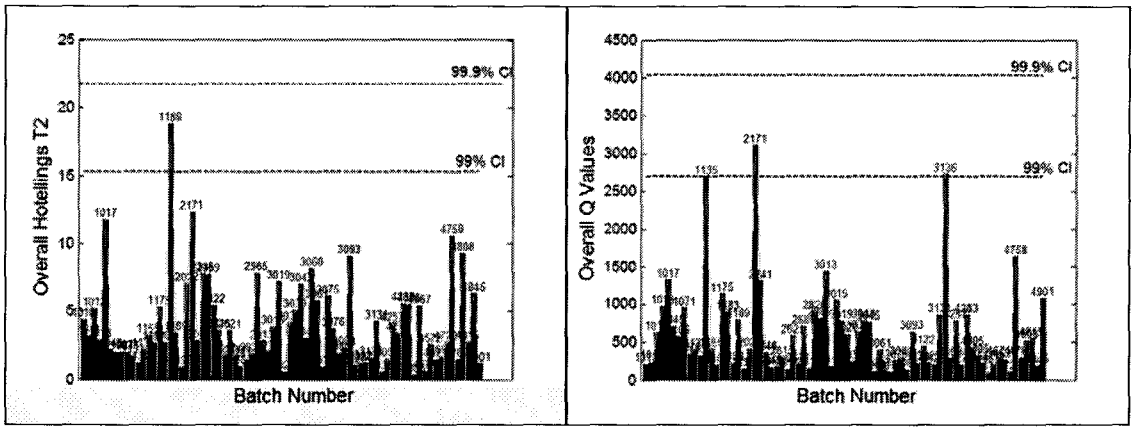

Figure 4.10 Overall Hotelling's T^2 and Q value plots for the reference data set.

4.6 MPCA Monitoring Scheme Performance

This section aims at evaluating the performance of the final MPCA monitoring scheme with relation to two aspects: number of false alarms and fault detection capabilities. This evaluation is performed in an off-line manner using all "in-control" and "out-of-control" test data sets.

Additionally, the applicability of different fault identification methods are discussed.

4.6.1 False Alarms

As previously mentioned in subsection 4.4.2, it is important for any monitoring system not to present an excessive number of false alarms so as not to affect operator acceptability. In order to verify if the final MPCA model meets this requirement, an off-line analysis is performed on a test set of 17 "in-control" batch data. None of these batches were previously used in building the reverence model. The test data set is composed of batches with different characteristics: i) 2 batches randomly excluded from the database which was later used to build the reference model; ii) 8 batches produced

around 4 months after those pertaining to the reference database; iii) 7 batches in which only 1 or 4 coils were annealed.

In order to further desensitize the final model with relation to Type I errors, an alarm is only generated after three consecutive SPE values are above the 99.9% C.I. or one Hotelling's T^2 value ia above the 99% C.I.

Of all "test" batches passed through the monitoring procedure, only 4 generated an "out-of-control" signal. Two of these batches were considered as "fast" batches while the other two had slightly different behavior at the end portion of the T1 profile. These alarms can easily be eliminated from a future model through the addition of more batches that present these particular characteristics. This is characterizes the iterative nature of model building; the reference dataset can be repeatedly augmented until unwanted alarms are eliminated.

From these results it is concluded that the final MPCA model does not present an excessive number of false alarms. Annealing plant personnel were satisfied with the MPCA model performance with relation to this metric.

Additionally, it is verified that this model is capable of handling all coil staking configurations used and also that the plant in question does not present significant drifts in operating conditions through the period of a few months. In case drifts in operating conditions had been detected, the application of an adaptive MPCA model could have been attempted.

4.6.2 Fault Detection

Although control algorithms are used in order to perform set point tracking of selected temperature trajectories within the annealing process, each batch run is subjected to the possible occurrence of five main fault types:

- Type 1 faults – Caused by a severe form of abnormality in furnace operation, as a consequence T1 values drop at a fast rate, followed by all other temperatures.
- Type 2 faults – Caused by a second form of abnormality in furnace operation. If this occurs at the beginning of the batch, T1 is slow at reaching it's set point (and all other temperatures are sluggish in rising as well). If this happens later on in the batch, lower than normal T2 values are registered.
- Type 3 faults – Fault caused by an equipment failure within the protective cover. This impacts convective heat transfer and is characterized by higher than normal T2 and low T3 values.
- Type 4 faults – The cause of this fault is normally attributed to a sticky valve in the furnace fuel inflow line and is characterized by fluctuation in T1values during soaking.
- Type 5 faults – The cause of this fault is not determined; one hypothesis is that this is the consequence of operator corrective actions prior to the development of faults types 1 or 2. This fault is characterized by short-lived drops in T1.

Currently at the batch annealing plant under study, detection of type 1 faults is done through the use of an on-line sensor. Detection of all other faults is carried out though a set of univariate, data based, check points that verify if the shape and values of the temperature trajectories are as expected. Both fault detection methods are subjected to failure and thus the operators are heavily relied upon to visually inspect the temperature trajectories of evolving batches.

In order to evaluate the performance of the MPCA monitoring scheme with relation to fault detection, an off-line analysis is carried out using a test set composed of 25 "out-of-control" batches. All of the most common fault types to which the batch annealing process is subjected to are represented within this data set.

When passed through the monitoring procedure these faults lead to an alarm in the Hotelling's T^2 and SPE monitoring charts. The batch time at which these alarms occurred can be compared to those generated by the monitoring scheme currently installed at the annealing plant and, through visual inspection of temperature trajectories, to those when it is though that the fault began (Figure 4.11).

Visual inspection of Figure 4.11 indicates that, except for type 1 faults, the performance of the MPCA monitoring scheme is superior to that of the original (currently installed) monitoring scheme. Table 4.3 shows that this conclusion is representative of the entire test database.

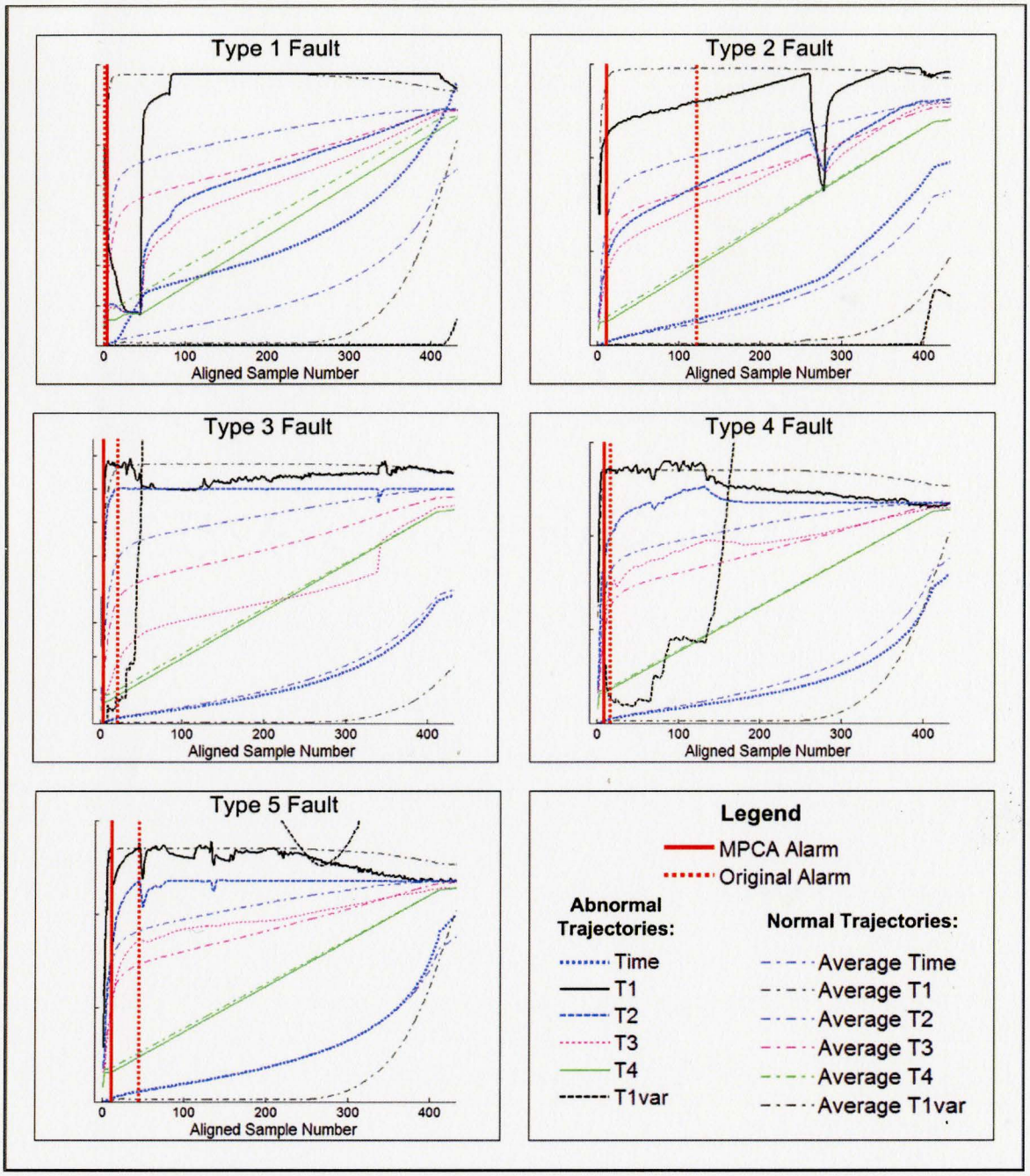

Figure 4.11 Temperature trajectories and alarms generated by the MPCA and original monitoring schemes for each main type of fault to which the annealing process is subjected to.

Table 4.3 Comparison between the MPCA and the original monitoring scheme.

Fault Type	Monitoring scheme with superior performance (faster fault detection)	Average time difference between the alarms (hours)
1	Original	0.22
2	MPCA	5.4
3	MPCA	0.22
4	MPCA	3.2
5	MPCA	0.22

Type 1 fault detection is currently performed by an on-line sensor and is thus always faster than any data-based approach. However, according to engineers working at the facility, this sensor is subjected to failure. In this scenario, the MPCA monitoring system is capable of signaling this fault within approximately 13 minutes of it's occurrence.

The gain in fault detection time by the MPCA model is very significant in the cases where fault types 2 and 4 occur (5.4 h and 3.2 h respectively). By detecting these faults sooner, faster corrective actions can be taken, total batch time reduced and final steel quality increased.

Additionally, the MPCA scheme was capable of identifying 4 out of 83 batches in which faults, not picked up by the original monitoring system, occurred. Two of these faults are shown in Figure 4.12. It is observed that the severity of these faults is significant and, thus, their detection is important.

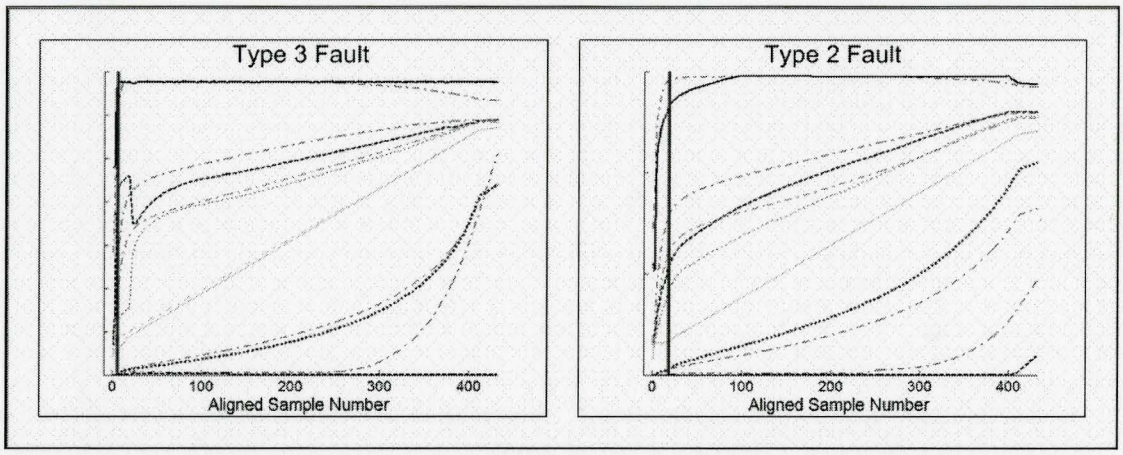

Figure 4.12 Temperature trajectories and alarm generated by the MPCA for faults which were not detected by the original monitoring scheme. See Figure 4.11 for the legend.

Only one batch processed under abnormal conditions did not cause the MPCA system to signal (Figure 4.13). This batch exhibits and unknown fault. The original monitoring scheme only accused a type 4 fault at the beginning of the batch. One way of ensuring that this type of fault can be captured in the MPCA system is to include the sum of the variances of T3 as an extra variable. Due to the low occurrence of this fault the addition of this extra variable was not considered necessary.

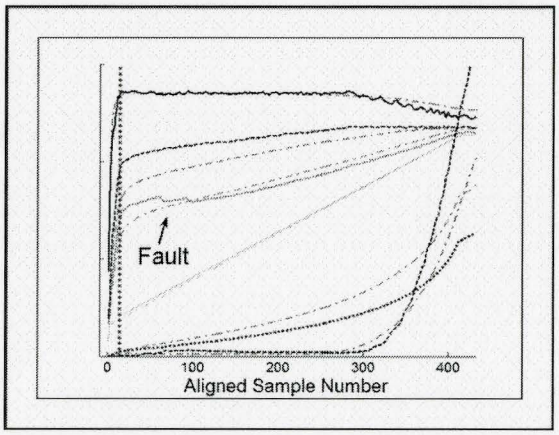

Figure 4.13 Temperature trajectories and alarm generated by the original monitoring scheme for an undiagnosed fault (not detected by the MPCA monitoring scheme). See Figure 4.11 for the legend.

Annealing plant personnel were very satisfied with the fault detection capabilities of the MPCA monitoring system.

4.6.3 Fault Diagnosis or Isolation

Posterior to fault detection, it is necessary to determine an assignable cause for the deviation. This is important so that appropriate actions can be taken to, either compensate for the fault in real-time, or to avoid future occurrences.

The most intuitive way of tackling this problem is to plot and visually inspect, in a univariate manner, all process variables pertaining to the anomalous batch. However, the number of variables is often very large, making it practically impossible to determine which subset of them is responsible for an "out-of-control" signal. Additionally, multivariate correlations between variables, which maybe responsible for the fault, are missed.

Kourti and MacGregor (1996) provide a review of various fault diagnosis procedures based on multivariate statistical methods. The most widely used (Qin, 2003) set of diagnostic tools are based on calculating the contribution that each variable has on individual scores (Miller *et al.*, 1998 and MacGregor *et al.*, 1994); as shown in Chapter 2.

However, very often, more than one score can present high values and individual investigation of every score plot can be fastidious. For these cases, Kourti and MacGregor (1996) suggested that an "overall average contribution" per variable be calculated:

$$CONT_j = \sum_{a=1}^{K} \frac{t_a^2}{s_a^2} p_{a,j}(x_j - \mu_j) \qquad (4.2)$$

where:

$t_{a,j}$ = score vector associated with the a^{th} principal component and the j^{th} variable;

$p_{a,j}$ = loading vector associated with the a^{th} principal component and the j^{th} variable;

s^2_a = variance of the a^{th} score;

x_j = process variable measurements;

μ_j = "in-control" population mean associated with the jth variable.

Additional rules suggested by Kourti and MacGregor (1996) to increase the discriminating ability of the overall average contribution plots are: i) only the normalized scores with high values ($K \leq n$) should be included; ii) set all negative contributions to individual scores to zero (i.e. the sign is opposite to the value of the score t_a).

All of the fault isolation methods mentioned in the previous paragraphs only cover the first of the two-step procedure involved in fault diagnosis: 1) find which variable(s) contribute to the "out-of-control" signal; 2) determine the root cause of the process upset. These methods rely on a causal relationship among the models and thus cannot provide direct fault isolation (Yoon and MacGregor, 2001). The second step is normally performed by a trained operators or engineer who use their process insight to provide feasible interpretations of the fault, on a case-by-case basis.

Automated fault isolation is possible by comparing signatures of current faults against a database of reference fault signatures. Current automated fault diagnosis methods differ in the type of signature used to characterize the faults and in the manner of comparing them against the reference signature bank. Qin (2003) provides an overview and analysis of statistical process monitoring methods for fault detection, isolation and reconstruction. If there are plenty of historical records with a wide variety of fault categories, classification and clustering methods are also available. Yoon and MacGregor (2001) propose an approach that extracts fault signatures that are vectors of movement of the fault in both the modeled and the residual space. Isolation is based on comparing the angles between the vectors of current and known faults.

However, due to their time-varying nature, application of automated fault isolation methods to batch data is not trivial. In order for these approaches to work, a fault library containing, not only all faults to which the process is subjected to, but also faults repeated at different moments of the batch, is necessary. This is due to the fact that faults affect the process differently when they occur at different parts of the batch. In some cases, not only the correlations between variables are different, but some variables may not even be measured at different batch sections. In the annealing process, the sum of T1 variability only starts once this variable has reached set point value. Thus, type 1 faults will have different signatures depending on if they occur at the initial or final sections of the batch.

In conclusion, for batch processes, reference fault signature banks can become prohibitively large. To the best of the author's knowledge, this issue has not been previously discussed in the literature.

Thus, in order to provide fault diagnosis for the annealing process, overall average contribution plots for Hotelling's T^2 (Kourti and MacGregor, 1996) and plain contribution plots for SPE values were used. These statistics, calculated at the sample in which the MPCA scheme first detected a fault, are shown in Figure 4.14 for each fault type.

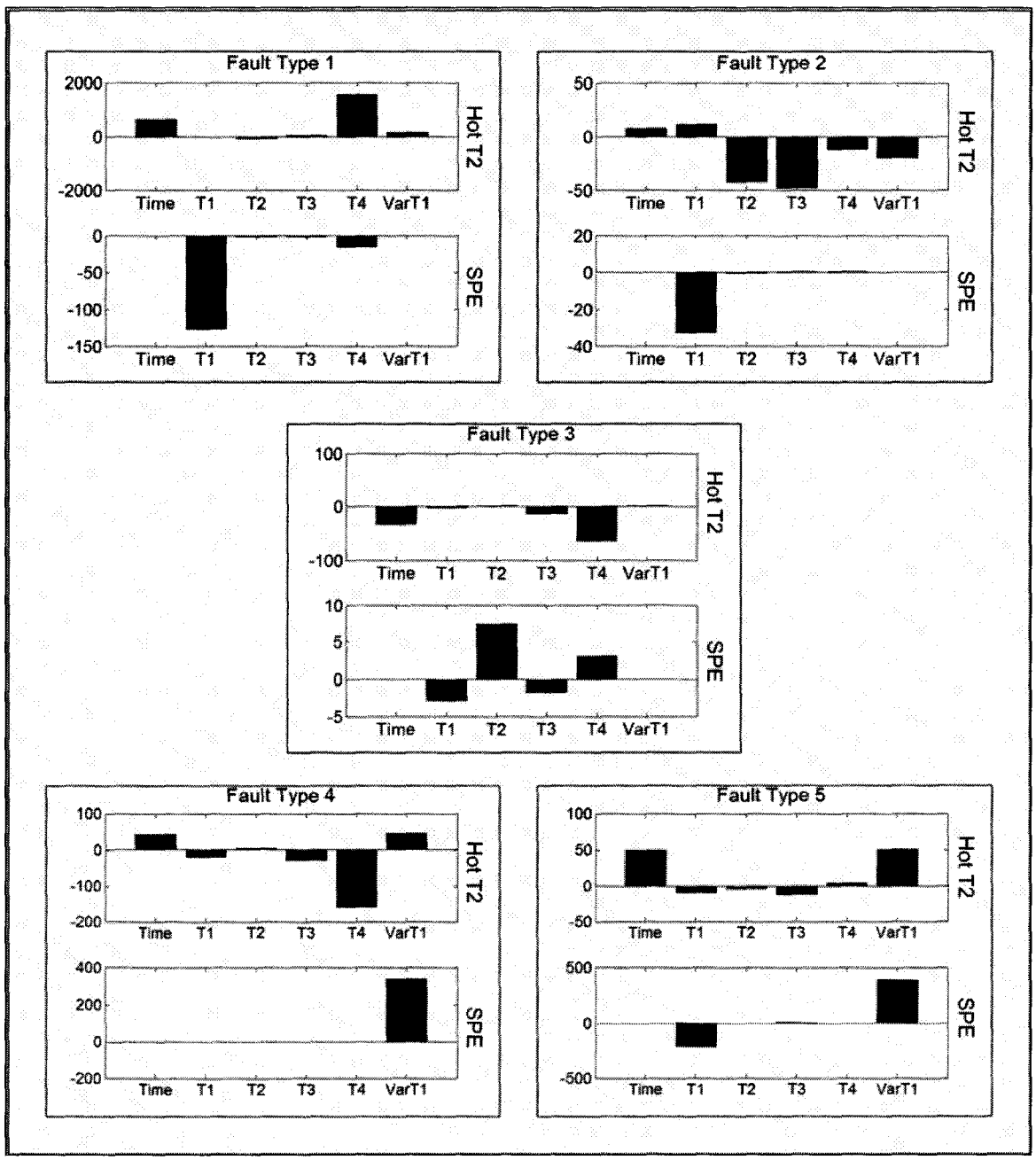

Figure 4.14 Overall average contribution plots for Hotelling's T^2 and plain contribution plots for SPE values were used.

Inspection of both the contribution plots to the Hotelling's T^2 and SPE, for all batches in the "out of control" test set, indicated that the contributions to the SPE statistics were much more informative at indicating the type of fault that occurred. Using

only the contribution to the SPE statistics it is verified that: i) fault types 1 and 2 both present lower than average values of T1; ii) fault type 3 presents higher than average values of T2; iii) fault types 4 and 5 present higher than average values of VarT1. These observations are in accordance with the theoretical descriptions of these faults given in 4.6.2.

It is not possible to distinguish faults 1 and 2 or 4 and 5 from each other due to the fact that, at their initial points, they progress in the same manner. Additionally, all type 1 faults, included in the "out of control" test set, occurred in the beginning of the batch however, if they were to occur at a later point, the SPE contribution plot would be similar to those for faults 4 and 5.

Even so, a decision-tree approach can be used to aid the operators or engineers in the fault identification process (Figure 4.15).

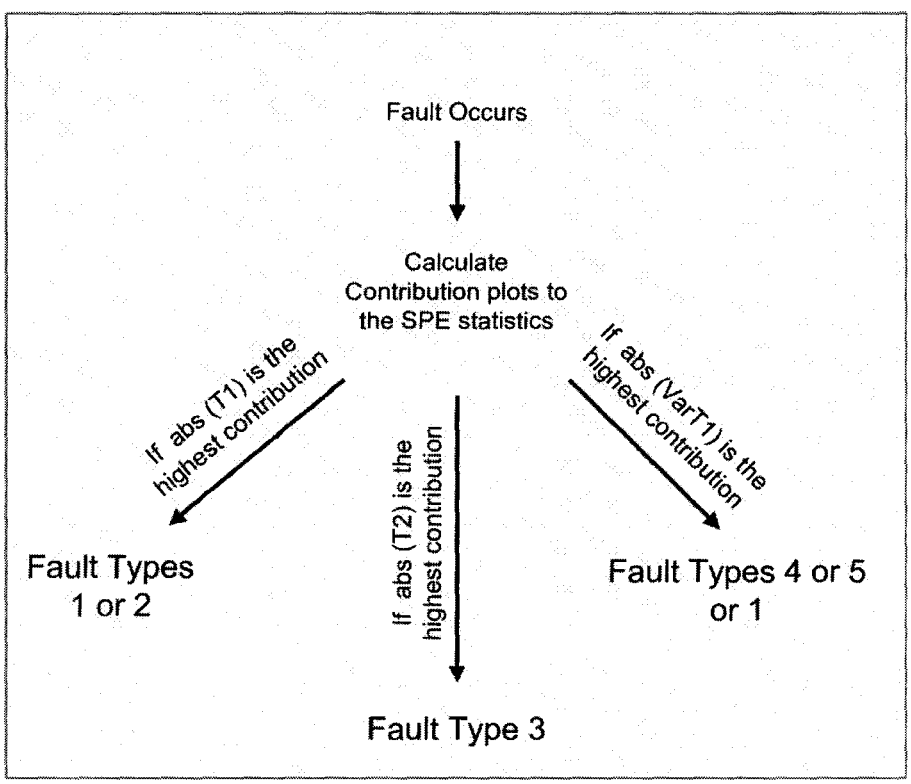

Figure 4.15 Decision-tree for annealing fault identification.

4.7 Conclusions

Multiway principal component analysis was successfully used to build a single, all-encompassing, on-line monitoring scheme for the heating phase of a multi-grade batch annealing process.

The performance of this system was evaluated based on pre-established false alarm and fault detection metrics and considered adequate for industrial needs. The MPCA monitoring scheme also presented superior fault detection performance when compared to the system that is currently in place at the annealing plant under study.

Issues relative to automated fault identification methods in batch process were also addressed within this work. Finally, a decision-tree based approach is suggested to aid in annealing fault isolation.

Further work includes extending the model to monitor the cooling phase of the annealing process.

Chapter 5

Pre-alignment of Batch Data for On-Line Monitoring

The purpose of the current Chapter is to present a simple alignment technique for batch data when on-line monitoring is intended. This technique relies on a PLS model, fitted to variables obtained prior to the beginning of each batch cycle (initial conditions), to predict the duration of new batches. The predicted time information is then used to set the sampling rate of in-coming process data.

This pre-alignment method is demonstrated on an on-line MPCA monitoring scheme built for the heating phase of a single-grade industrial batch annealing process. Additionally, various methods for dealing with matrices containing different sized observations, in this case resulting from the existence of two possible steel stacking configurations, are proposed and evaluated.

5.1 Process Description

Section 4.1 contains a description of the multi-grade batch annealing process under study. However, for the current Chapter, only one grade type and stacking configurations ranging from 2 to 3 coils are considered. In addition, prior to the beginning of each batch annealing cycle, the following variables are registered:

- SH and SV – Measured geometry variables for the steel stack that will be annealed.

- SDKF – Calculated variable which summarizes the geometry of the stack.
- CWTi and CTHKi – Measured weight and geometry variables (respectively) for each coil (i) within a stack.
- CKi – Calculated variable summarizing the geometry of each coil (i) within a stack.

5.2 Project Incentives and Objectives

Alignment or synchronization of batch data, necessary to achieve the assumptions of equal duration and matching trajectories inherent to MPCA and MPLS batch-wise unfolded models, is an issue still under discussion in the literature (Garcia-Munoz, 2004). While all of the alignment techniques listed in section 2.3 can be used for troubleshooting, when the purpose of the model is to monitor new batches, these options become restricted since the sampling rate must be set in advance.

Within this scenario, crude linear interpolation over the entire batch time or over specific stages (as shown in Chapter 3) and use of discrete events, cannot be employed.

While alignment for monitoring purposes might be easily achieved through the use of an indicator variable, often one cannot be found for every phases of the batch cycle. Even if a monotonically increasing variable is present, it's rate of increase may not be fast enough to allow the model to capture the variability of all other measurements. Additionally, whenever the indicator variable presents decreasing or missing values due to the occurrence of a process or sensor fault, monitoring is not possible. During the analysis of the batch annealing process data presented in Chapter 4, these practical issues were encountered and efforts had to be made to work around them.

When an indicator variable is not present in the data set, a simple alternative is to pre-specify the size of the training matrix according to either the shortest batch (data trimming technique) or a set of the longer batches present in the historical data-set (data augmentation techniques). In both cases, all in-coming data that exceeds the size of this

matrix is discarded and thus not monitored. Also, while these methods guarantee that the assumption of equal duration is met, no effort is made towards trajectory matching.

An evolving version of dynamic time warping (DTW) was introduced by Westerhuis *et al.* (2003) to allow for on-line alignment and monitoring of batch processes. Even though it is successful, the application of this technique is mathematically intensive.

The objective of the current work is to present a simple alignment technique for batch data when monitoring is intended. The method here proposed is based on a suggestion made by Kourti (2003) for predicting batch lengths for those cases in which this variable is a function of only one other variable, which is known *a priori* (recipe, grade, etc.). The technique here considered, referred to as pre-alignment, relies on a PLS model, fitted to a set of historical data collected prior to the beginning of all batch runs (initial conditions), to predict the duration of new batches. The predicted time information is then used to set the sampling rate of the in-coming data (time divided by expected length).

However, due to prediction errors, not all batches will automatically have the same number of sample points. Data trimming or augmentation techniques must also be used for this to be achieved. The advantage is that, the pre-aligned data have observations with more similar number of sample points then the raw data, thus leading to a reduction in samples that are discarded as a consequence of using complementary alignment techniques. More importantly still is that, by pre-aligning the data, trajectory matching is also being performed. The question of if the level of alignment achieved is satisfactory depends on the nature of the batch trajectories. This objective will be achieved for those sets of data in which a linear expansion or compression over the entire batch trajectory leads to synchronization. In all other cases an improvement in synchronization maybe achieved by compressing or expanding only a select number of stages.

The pre-alignment technique proposed is demonstrated on an on-line MPCA monitoring scheme built for the heating phase of a single-grade industrial batch annealing process. Processes which are highly dependent on one or more cooling stages are also

good candidates for pre-alignment since the capacity of this utility, and thus the time it takes to achieve temperature set points, is normally seasonal and reliant on overall plant demand at a given time. Another suggested application for this technique is for cases when batch run durations depend on achieving final quality specifications that are a function of known raw material quality.

5.3 Description of the Data Set

The historical data set used for the current study included a total of 35 "in-control" and 6 "out-of-control" batch runs. Each batch run contains:

a) 9 to 12 variables (depending on the number of coils being annealed) reflecting the geometry of the steel stack and individual coils. These variables are individually detailed in section 5.1.

b) 4 variables reflecting process equipment configurations - base location (Pn), furnace type (FT), furnace number (FN) and fan speed (FS) - that will be used to anneal a specific steel stack. Variables Pn and FT are qualitative variables with six and two levels, respectively. These were substituted for a total of seven new variables to which values of either zero or one were assigned (Montgomery and Runger, 1994).

c) 4 process variables (T1, T2, T3 and T4) and 2 calculated variables which reflect cumulative time per batch (Time) and the variance of T1 (T1Var).

The variables described in *a* and *b* are available prior to the beginning of each batch run and are included in the Z-matrix (initial conditions). The variables described in *c* are sampled at 5-minute intervals during the heating phase of the process and are, subsequent to alignment, included in the X-matrix (process variables).

Additionally, in order to build the PLS predictive model, the total processing time for the annealing heating phase of each batch run was registered in the Y-matrix (quality variable).

Throughout this work various software were used: SIMCA-P+ version 10.0 developed by Umetrics (sections 5.4.1-5.4.4); Batch SPC developed by McMaster University (section 5.5) and JYPLS developed by Munoz (section 5.4.5).

5.4 Evaluation of PLS Predictive Models

Batch data pre-alignment, as it is proposed, initiates with the use of a statistical model to obtain a prediction of the duration of new batches.

With respect to the annealing process, a predictive model can be obtained by fitting the data available prior to the beginning of each annealing cycle (Z-matrix) to the heating cycle duration (Y-matrix) for a set of reference batch runs. Due to the multivariate and highly correlated nature of this data set, PLS is the statistical modeling method of choice. However, since the steel stacks can have either 2 or 3 coils, the complete Z-matrix is composed of observations (rows) of either 18 or 21 variables (Figure 5.1). This poses an additional challenge for building the predictive model. The current section explores and evaluates various techniques that, when combined with PLS, can be used to overcome this issue.

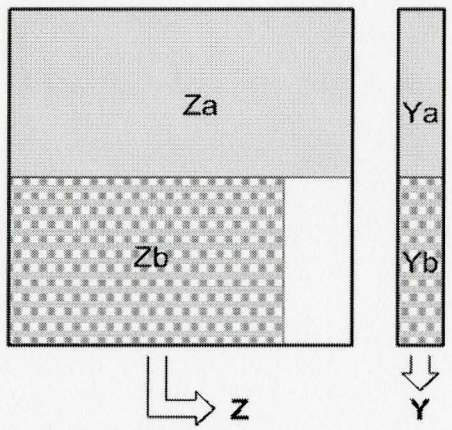

Figure 5.1 Representation of the complete Z and Y matrices, matrices containing only data relative to steel stacks with 3 coils (Za and Ya) and matrices containing only data relative to stacks with 2 coils (Zb and Yb).

5.4.1 Separate PLS Models

From a latent variable modeling perspective, the most natural approach to handling the four data matrices available (Za, Ya, Zb and Yb) is to fit two separate PLS models, one for each Z-Y matrix pair (Munoz *et al.*, 2005). Cross-validation indicated that 2 PCs were optimal to parsimoniously explain (R^2Y) and predict (Q^2) the total variation in both Ya and Yb. An overview of these models is shown in Table 5.1.

It should be noted that two observations were excluded from the Zb-Yb PLS model due to high leverage and distance from the model.

Table 5.1 Overview of the PLS models for data sets with 3 (Za-Ya) and 2 coils (Zb-Yb).

Data Set Modeled PLS	R^2Y (%)	Q^2Y (%)
Za-Ya	84.4	70.9
Zb-Yb	92.9	76.3

Inspection of Table 5.1 shows that, in both cases, the batch-to-batch variation in annealing cycle time is well explained and predicted by process equipment configurations and physical and geometrical measurements of the steel stack and coils. It is of interest at this point to determine which of these variables have the highest impact on Ya and Yb.

Plots of PLS weights and regression coefficients (B) are useful tools for determining the relationship between factors and responses. Regression coefficients are used to re-express the PLS solution through equation 5.1; equation 5.2 provides their relationship to PLS weights :

$$Y = BX + F \qquad (5.1)$$

$$B = W^*C \qquad (5.2)$$

In addition, a parameter called variable influence on projection (VIP) summarizes the importance of each factor, both for the Z- and Y-model parts, and aids in the interpretation of weight plots (Eriksson *et al.*, 1999). By definition VIP is a weighted sum

of squared PLS weights. Predictors with VIP values significantly larger than 1 are considered as being most influential on the model.

Simultaneous inspection of the plots shown in Figures 5.2, 5.3 and 5.4 indicates that variables CK and CWT relative to coils 1, 2 and 3 are all positively correlated with Y and important predictors of this quality variable. Variable SV is also of some importance.

Physically these results show that the heavier the coil, the longer it takes T4 to reach set-point value and thus, the more time-consuming the heating phase of the annealing process is.

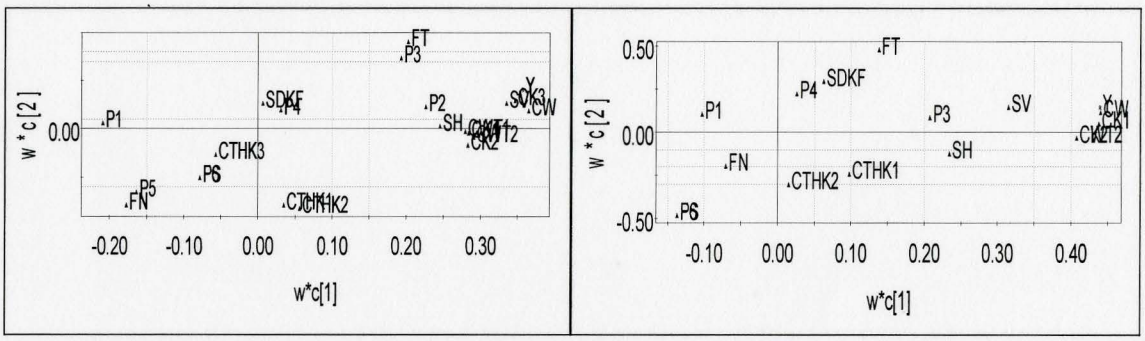

Figure 5.2 PLS weights for the data sets with 3 (left) and 2 (right) coils.

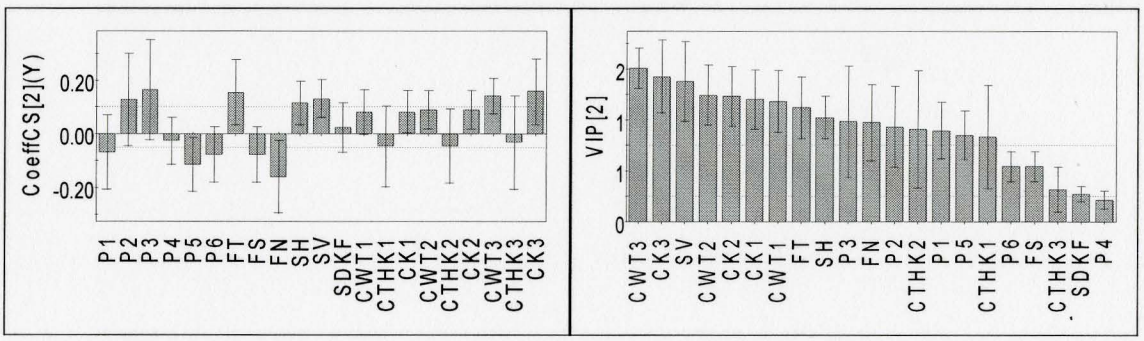

Figure 5.3 PLS regression coefficients (right) and VIP plots (left) for the data set with 3 coils.

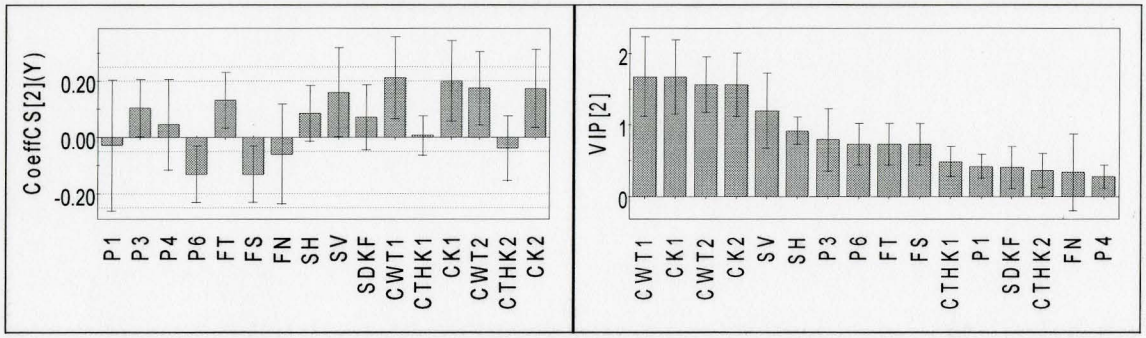

Figure 5.4 PLS regression coefficients (right) and VIP plots (left) for the data set with 2 coils.

With the intent of increasing model performance and decreasing modeling efforts, the next three subsections focus on evaluating if a single PLS model can be used to predict batch cycle durations.

5.4.2 Physical Parameter Value Substitution

The first approach under evaluation for handling the different sized observations in the Z-matrix with a single model consists of considering that, for all stacks with only two coils, the third coil can be represented as having null mass and dimensions. In this manner, variables CWT3 and CTHK3, which are originally inexistent for stacks with only two coils, are set to zero. Calculated variable CK3, which is a function of both CWT3 and CTHK3 and tends to infinity when these tend to zero, is set to a high random number (100).

PLS-modeling of this new Z-matrix and the original Y-matrix, both of which contain all the observations within the historical data set, yielded a two-component model capable of explaining 84.5% (R^2Y) and predicting 65.8% (Q^2) of the total variation of Y. No observations were excluded.

Simultaneous inspection of weights, regression coefficients and VIP plots for the PLS model in question (Figures 5.5 and 5.6), indicates that only variables CK and CWT relative to coils 1 and 2 (and, to some extent, SV), are positively correlated with and

important predictors of Y. However, due to the similar nature between all CK and CWT variables and the results obtained in section 5.4.1, it is also expected that CK3 and CWT3 have an analogous impact on Y.

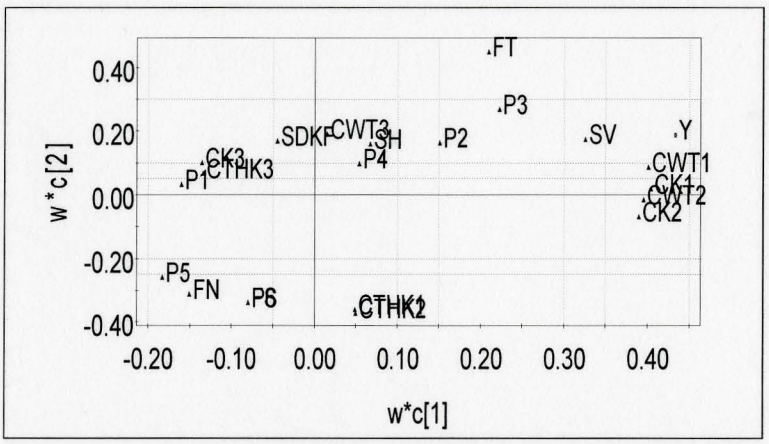

Figure 5.5 PLS weights.

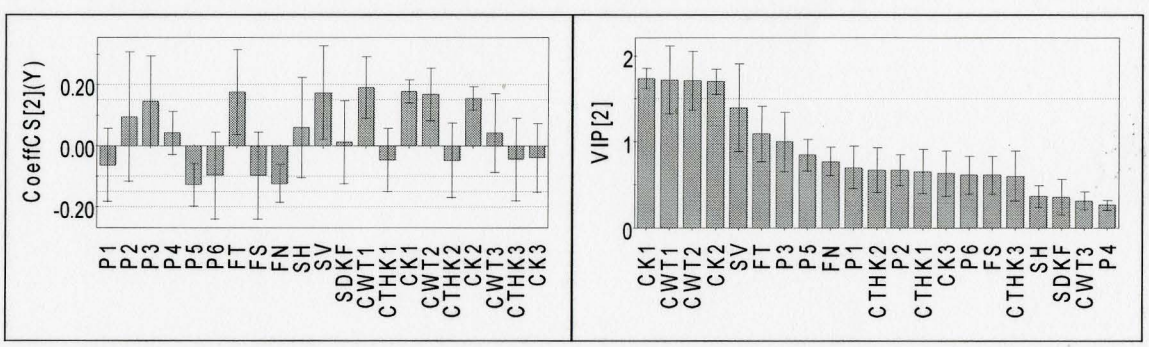

Figure 5.6 PLS regression coefficients (right) and VIP plots (left).

By plotting CK and CWT *versus* cycle duration for all existing coils in a univariate manner (Figure 5.7), it is observed that, in fact, all these factors are equally important in explaining and predicting the variations in Y. Simple linear regression models indicated R^2 values ranging from 26% to 46% for all coils.

It is therefore concluded that the PLS model in question is not capable of correctly identifying all the predictors of the batch annealing heating cycle duration (Y). This occurred because, even though it makes physical sense to do so, by setting the

measurements of inexistent coils to 0 or 100, variables CK3, CTHK3 and CWT3 remain constant regardless of the value Y assumes. Thus an unwanted break in the correlation occurs, causing the variable substitution method attempted to have non-satisfactory performance.

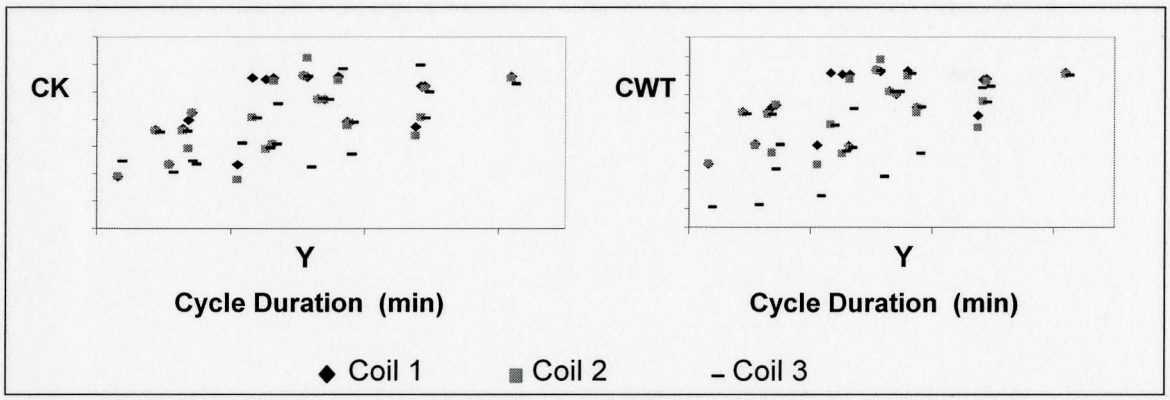

Figure 5.7 Plots of CK (left) and CWT (right) *versus* Y for all existing coils.

5.4.3 Missing Data Substitution

One of the features that motivates the use of PLS for modeling process data is it's capability of handling missing information. Several latent variable techniques for estimating scores from data with missing measurements are presented by Nelson *et al.* (1996). Of the missing data methods, the one most widely used (and applied throughout this work) is based on the NIPALS algorithm and is known as the single component projection method (SCP). This technique was first proposed by Wold in 1964 (Nelson *et al.*, 1996) and within it, all iterative regressions necessary to determine PCA or PLS model parameters (scores, loadings and weights) are performed using only the data that is present and ignoring the missing points. This is equivalent to either: i) setting the residuals for all missing elements in the least squares function to zero in each iteration; ii) replacing the missing values by their minimum distance projections onto the current estimate of the loading or score vector at each iteration (Nelson *et al.*, 1996).

A rule of thumb given by Eriksson *et al.* (1999) for the relative amount of missing data that can be handled by the NIPALS algorithm is 10% to 20% for an ordinarily sized matrix of 50-100 observations. According to Nelson *et al.* (1996), the matrix should have 5 times as many observations in any row or column as the number of dimensions being calculated. However, it is important to remember that these are crude approximations; the amount of missing data one can have depends on the correlation patterns of the variable which contains the missing values relative to other variables in the data set. Additionally, data should not be missing according to a systematic pattern.

The technique proposed in this section with the purpose of handling the unequal sized observations that compose the Z-matrix, consists of setting all measurements relative to inexistent coils as missing data. The resulting Z matrix contains 5% of missing data (overall), never exceeding 28% on a single column or 15% on a single row. These values are slightly above those recommended by the literature, thus requiring that the performance of the final PLS model be closely evaluated.

By fitting a PLS model to the Z-Y matrices containing the missing data, it is verified that this model is capable of explaining 86.0% of the variance in the quality data (R^2Y) and predicting 75.3% of it's total variation (Q^2) with two principal components. No observations were excluded.

Simultaneous inspection of the plots shown in Figures 5.8 and 5.9 indicates that variables CK and CWT relative to coils 1, 2 and 3 and SV are all positively correlated with Y and important predictors of this quality variable. This result is in accordance with those obtained in section 5.4.1 and the observations made in section 5.4.2.

It is thus concluded that the technique proposed in this section is capable of handling unequal sized observations with superior predictive power relative to all other methods considered. This is very impressive and indicates that missing data substitution is very useful when smaller, different sized and correlated, data sets are involved.

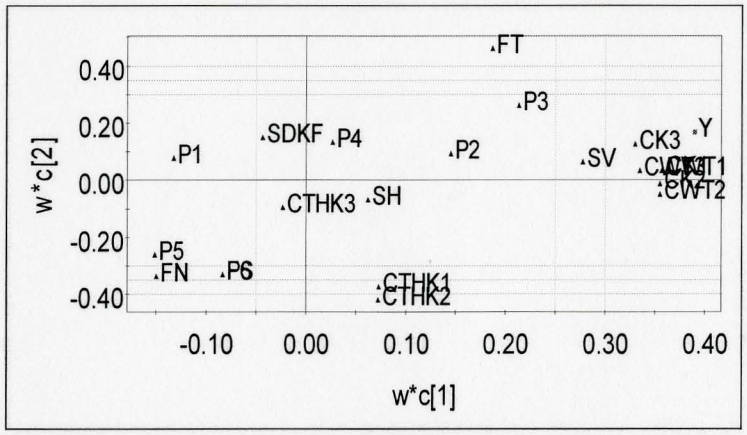

Figure 5.8 PLS weights.

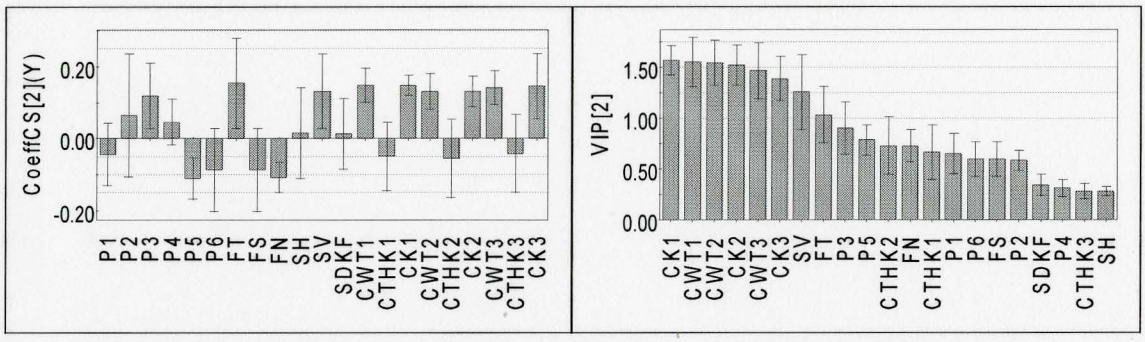

Figure 5.9 PLS regression coefficients (right) and VIP plots (left).

5.4.4 Joint-Y PLS

Joint-Y PLS (JYPLS) is a latent variable regression method proposed by Munoz *et al.* (2005) with the purpose of modeling the common latent variable structure in multiple plants. This technique was initially conceived to solve product transfer problems and, subsequently, a suggestion was made to apply it to parallel plant assessments. The purpose of the current section is to extend the use of JYPLS to handle matrices with unequal sized observations.

The basic concept behind JYPLS is that Ya and Yb lie in a common latent variable plane and thus, can be jointly defined by a common loading matrix Q_j if a single PLS model is built. While no size restrictions are imposed on the X-matrix (or Z-matrix,

according to the notation used for this chapter), JYPLS requires that both Y matrices have the same number of variables (columns). In their work, Munoz *et al.* (2005), also provide an overview of the diagnostic tools that can be used to assess a JYPLS model. According to these authors, basically, the only difference in the calculation of these values from separate PLS models is that the residuals for both of the Y matrices are computed using the same Q_j loading matrix.

The JYPLS model, fitted to the Za-Ya and Zb-Yb matrices, has overall R^2Y and Q^2 values of 82.5% and 60.0%, respectively. These values are somewhat lower than those obtained by the methods proposed in the previous sections. This disparity can be attributed to the fact that JYPLS is a more general algorithm. This method assumes that not all variables within Zb are the same as in Za; it considers that purposeful changes are made in operating conditions so that a difference in grades can be achieved (Munoz et al., 2005). This differs from the missing data approach, wherein it is assumed that all "inexistent" data are actually missing values of the variables previously modeled; the correlation structure is thus defined by common cause variation. Since, for the annealing case, all missing data can be correctly described by the variables of existing third coils, incorporation of this previous knowledge leads to better results in terms of predictability.

Comparing Figure 5.10 with 5.2 shows that JYPLS is capable of correctly identifying the predictors of both Ya and Yb. These are promising results and motivate future use and studies of JYPLS.

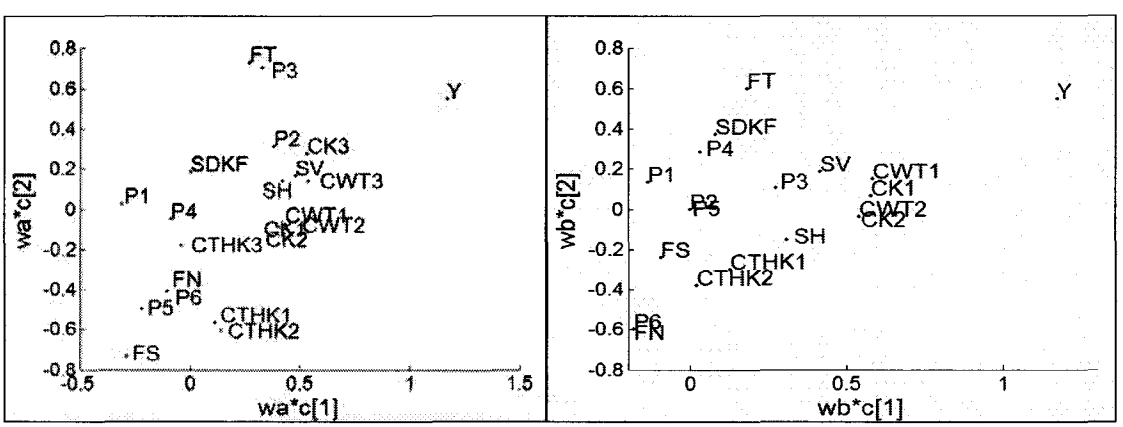

Figure 5.10 PLS weights for Za-Ya (left) and Zb-Yb (right).

5.5 MPCA Monitoring Scheme

Since the purpose of pre-alignment is synchronization of batch data when on-line monitoring is intended, it is important to compare the performance of a monitoring model built and used with data pre-processed in this manner with that of models built using other alignment techniques. The following alignment techniques were chosen for comparative purposes: i) trimming of the data set based on the batch of shortest duration; ii) usage of an indicator variable with the objective of trajectory re-sampling; iii) crude linear interpolation over the entire batch time. Among all of these, crude linear interpolation is the only one that cannot be applied for on-line monitoring. It's purpose within this study is to illustrate the impact of the predictive error inherent to the pre-aligning technique used.

With the intent of allowing for a fair comparison between the methods, data trimming was also applied to the pre-aligned data based on the batch of shortest duration. Since only one grade type is present within this data set, no grade-specific mean centering was necessary. Fault type identification and control limits used for detection are the same as those described in Chapter 4. The results of the comparative study between monitoring models built using the different alignment techniques described are shown in Table 5.2.

Additionally, Figure 5.11 shows the results of the application of these techniques for variables T4 and Time. Note how information regarding time usage throughout each batch run is successfully kept by variable Time when pre-alignment, crude linear interpolation and indicator variable methods are used.

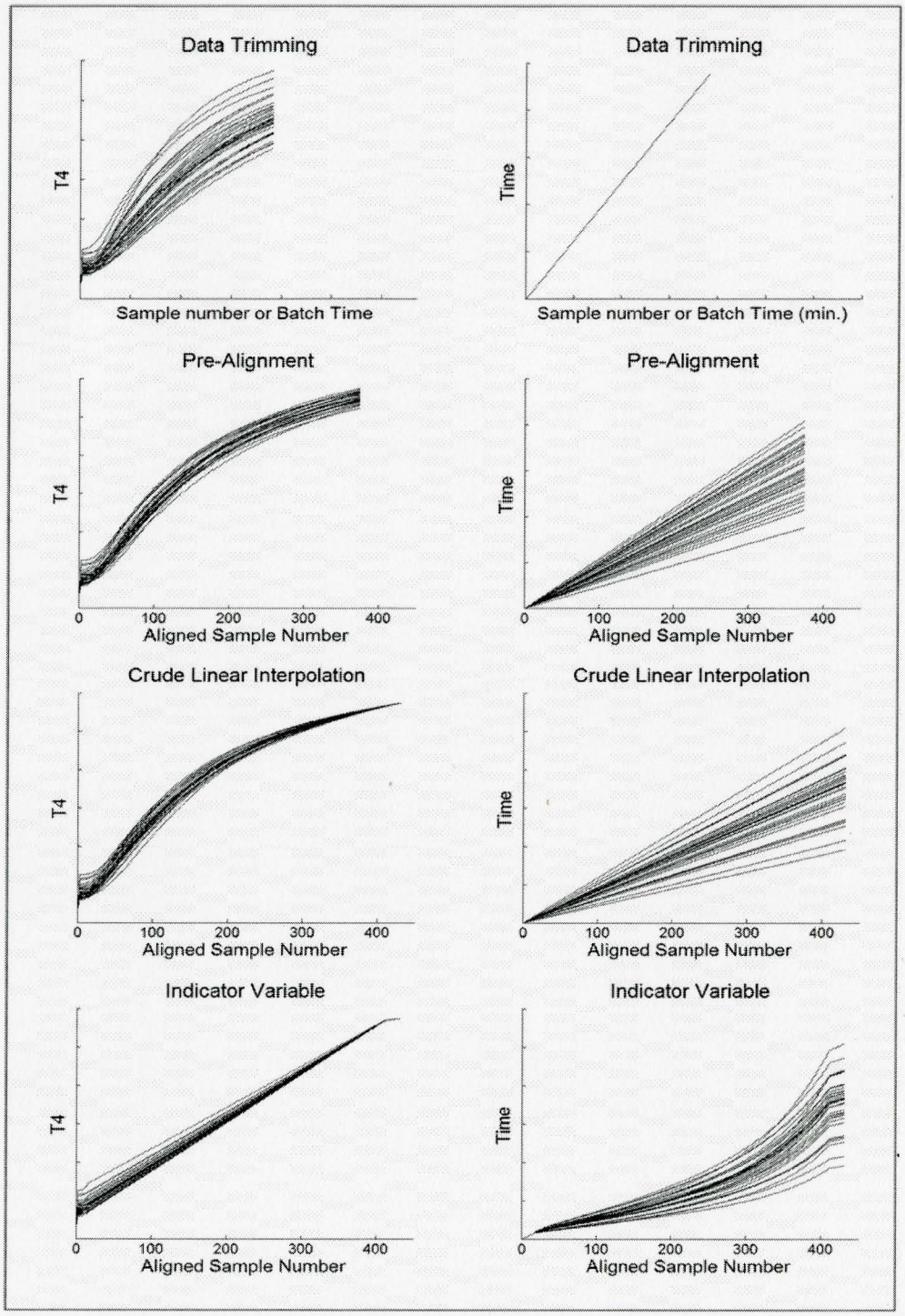

Figure 5.11 Profiles for variables T4 and Time obtained from aligning the batch annealing data set using (from top to bottom): data trimming, pre-alignment, crude linear interpolation and indicator variable methods.

Table 5.2 Results of the comparative study between alignment techniques.

Alignment Technique	**Pre-Alignment**	**Data Trimming**	**Crude Linear Interpolation**	**Indicator Variable**
MPCA model results (%)				
R^2X	78.8	84.6	82.7	80.8
Time of fault identification (min)				
Fault Type 1	45	55	50	50
Fault Type 2	135	135	135	130
Fault Type 3	30	40	40	40
Fault Type 4	200	225	200	215
Fault Type 5	210	215	200	210
Fault Type 5	200	210	200	215
Mean batch time not monitored due to data trimming (hours)				
	3	14	0	0

The data presented in Table 5.2 shows that the MPCA model built using the data trimming method of alignment is capable of explaining the highest amount of variability in the X-matrix. This is due to the fact that such a large amount of data had to be excluded from the model due to the limitation of the batch of shortest duration, leaving less variation in trajectory profiles to be modeled. The mean amount of time that is not monitored per batch due to the use of this technique is 13 hours or 32% of the total batch duration, value much larger than the 3 hours or 7.5% of the total batch duration not monitored by pre-alignment. Additionally, monitoring using pre-aligned data showed consistently superior performance with relation to fault identification when compared to data trimming.

Pre-alignment showed a comparable performance relative to fault detection when compared with linear interpolation, indicating that the error in batch time prediction was not of practical consequence. With relation to the use of an indicator variable, pre-alignment showed a slightly superior performance, mostly likely due to the fact that the

rate of increase in the indicator variable was not capable of capturing as much variability in other measurements, specially during the final few hours of the run.

More significant deviations among these performance results might be verified if this comparison study is applied to a data set with larger batch-to-batch variations in terms of alignment.

5.6 Conclusions

In this Chapter, a data pre-alignment method was proposed and successfully used to synchronize batch data for on-line monitoring purposes. The performance of this technique relative to other, more traditional alignment methods, was evaluated. For this purpose, an on-line MPCA monitoring scheme, built for the heating phase of a single-grade industrial batch annealing process, was used. The results of this comparative study indicated that data pre-alignment has a consistently superior performance relative to data trimming and an equivalent performance relative to crude linear interpolation and the use of an indicator variable.

Various methods for dealing with matrices composed of different sized observations were also proposed and evaluated. The method which presented best overall performance (highest predictive and predictor identification abilities) consisted of a single PLS model fitted to a data set in which measurements relative to inexistent coils were set as missing.

Chapter 6

End-Point Prediction of an Industrial Batch Process

The main purpose of this chapter is to illustrate the use of the multivariate latent variable methodology proposed by Nomikos and MacGregor (1994, 1995a and 1995b), together with a data pre-processing technique suggested by Marjanovic *et al.* (2006) , in order to predict the end-point of an industrial batch process. Results of this study are presented and assessed based on the types of process variables included in the data set.

Additionally, the importance of identifying product quality variables that correctly reflect those required by it's final application is exemplified.

6.1 Process Description

The process under study is composed of a batch reactor, a neutralization tank and a storage tank. Among other products, this unit synthesizes a polymeric additive which is used as a dispersant for cement or concrete admixture. This product is obtained through a three stage process: monomer production, condensation polymerization and neutralization, respectively.

During the first stage, reagents 1 and 2 are added to the reactor, in a semi-batch manner, and the mixture is heated to a pre-specified set point value for a fixed number of hours. Once this point has been reached, the reaction is cut with quench water. In sequence, a third reagent is added and the temperature again raised to allow for polymer formation through condensation reactions (stage 2). This stage is again terminated after a

pre-specified number of hours through the addition of more quench water. Constant agitation is performed during the whole process.

Figure 6.1 depicts typical bulk temperature behavior (within the batch reactor) and phase indications throughout a batch run.

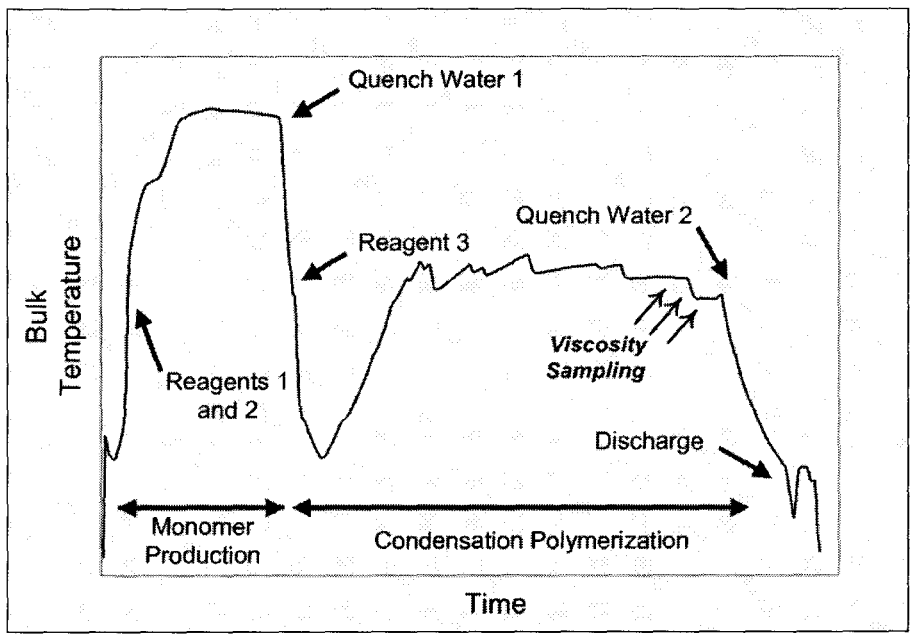

Figure 6.1 Typical bulk temperature behavior and phase indications for the production of the polymeric additive under study.

After being cooled and discharged from the reactor, the synthesized material is transferred to the neutralization tank where a basic solution is added. Samples are taken at the end of this stage to verify if product pH and two other physical characteristics are within specification. If this is the case, the final product is transferred to a storage tank.

Recorded process variables throughout stages 1 and 2 are listed bellow:

- Temp – Bulk Temperature in the reactor;
- Pres – Pressure in the reactor;
- R1W, R2W, R3W and WaterW– Cumulative weight of Reagents 1, 2 and 3 and quench waters;

- R1F, R2F, R3F and Water F –Flow rates for Reagents 1, 2 and 3 and quench waters;
- Coolant Water – Coolant water flow through the reactor jacket;
- CV1 and Temp CV 2 – Control valve readings, reflecting the alignment of either coolant water or vapor into the reactor jacket;
- Vent Valve – Variable showing if the reactor vent is open or closed;
- Agitator – Variable showing if the agitator is on or off;
- Valve – Variable showing if the valve with liberates the addition of Reagent 2 is open or closed;

Posterior to a revamp of the unit under study, in which the original reactor was substituted for one with more than three times it's volume, customers identified a significant reduction in the overall performance of the polymeric additive and an increase in it's variability. Product performance is quantified by the customers through a slump test. This type of test reflects the fluidity/workability of the concrete and basically consists of: i) preparing a concrete mixture containing a fixed amount of the additive under study; ii) placing the mixture in a mold; iii) withdrawing the mold; iv) measuring the height of the remaining concrete pile and subtracting this value from the height of the mold. In general, high slump values are indicative of good product performance; however, a lower bound for this measurement is also established. Further description of this test is given in Ramachandran (1997).

Even though a loss in product performance was verified by customers through slump test values, no changes in measured process or laboratory tests were verified by the production plant personnel.

6.2 Project Objectives

Project objectives are two-fold: initially the identification of a product quality variable which impacts slump test results, is required; in sequence, a control methodology

for this variable must be implemented. The control strategy should be possible to implement in the plant environment, require as little effort and be as safe as possible.

6.3 Quality Variable Identification

Literature research indicated that the molecular weight of a plasticizer is an important factor in determining the viscosity (and thus the slump value) of the cement paste or concrete mixture to which it is added. However, this only occurs up to a certain point, after which, any further increase in polymer molecular weight does not effect it's performance. Additionally, it was reported that polymeric additive producers often do not directly measure this variable due to it's inherent difficulties. These issues are clearly discussed in Ramachandran (1997) and Aitcin (2000), along with the information that polymer viscosity reflects an increase in it's molecular weight.

Based on these facts, a control methodology was proposed in which simple Brookfield viscosity measurements of polymer samples taken at the end of the condensation phase, would be used to determine when the batch should be terminated. This method is also based on the prior knowledge that polymer molecular weight increases as the condensation reaction progresses in time.

In order to test if this method would work and to determine an optimal target viscosity value, a bench-scale test was employed. This test is known as minislump and it holds a direct correlation to slump test results (Ramachandran, 1997). Minislump test procedures adopted throughout this work were based on those described by Monte (2003).

Polymer samples were collected at various points throughout the condensation phase of the additive's production and their viscosity and minislump values were determined. The results of this study are shown in Figure 5.2 (left). It is important to note that, due to the fact that these samples were taken prior to the neutralization phase, they had to be neutralized in the laboratory prior to the minislump tests.

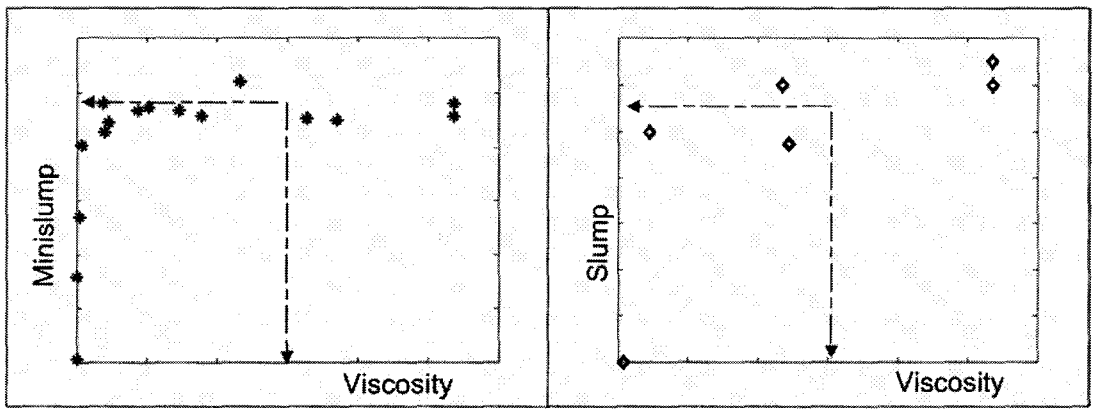

Figure 6.2 Correlation between the additive's acid phase viscosity and final product minislump (left) and slump (right) values. Optimal end-point viscosity values are also indicated.

Visual inspection of Figure 6.2 (left) indicates that, in fact, minislump values increase with polymer viscosity up to a certain point. This point represents that which the condensation reaction should be stopped, since additional processing time will not impact product performance but will impact production costs. The target viscosity value was thus determined by cross-examining the minislump results with those of a small number of samples sent to a customer so that their slump values could be determined.

In sequence, operators were trained to take hourly polymers samples until the desired final viscosity value is achieved in each batch run. When this point is reached, the reaction is terminated through the addition of quench water. In order to decrease the number of samples, these start being collected only after a fixed number of hours has passed since the beginning of the condensation phase. This is indicated in Figure 6.1.

Figure 6.3 shows the results of controlling the batch end-point through viscosity measurements. This graph contains the slump values measured by a customer whenever a new production lot of the additive was received, over several months. The upper and lower limits indicated (U.L and L.L) were set by the customer. Lots falling outside this specification could potentially be returned.

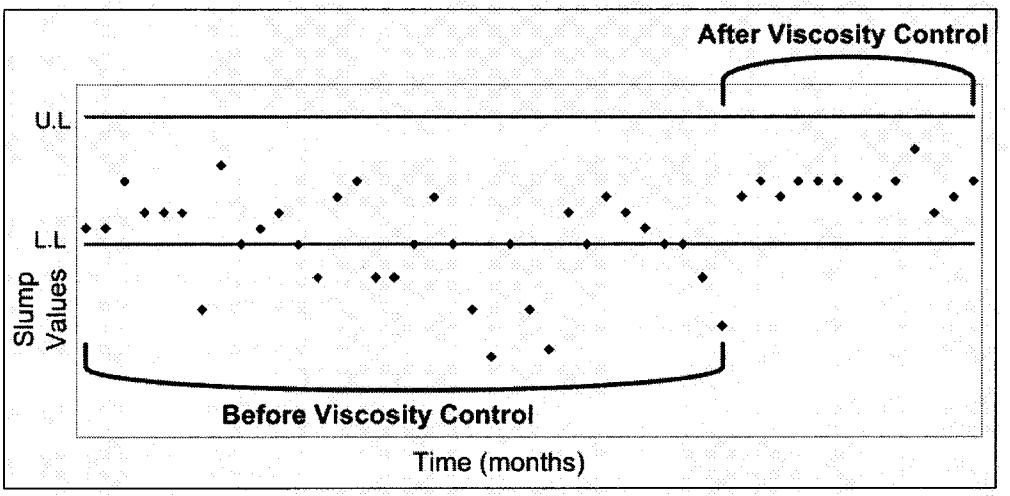

Figure 6.3 Shewhart chart showing the slump values of the lots of additive received by a customer prior and posterior to the implementation of the viscosity-based end-point control.

Inspection of Figure 6.3 indicates that the control methodology proposed was successful at specifying the additive in terms of it's final application. This shows the importance of identifying final product quality variables that correctly reflect those required by the customer's application.

In spite of the success achieved in terms of quality control, the sampling technique used is very time consuming and involves operational and personal risks. Additionally, laboratory analysis of the samples usually take anywhere from 30 minutes to an hour. This occasionally leads to unnecessary reaction times and thus, non-optimal conditions in terms of processing costs.

These factors serve as an incentive to attempt to use process data to build a soft-sensor capable of predicting polymer viscosity.

6.4 End-Point Prediction

This section describes the efforts made towards building a model capable of predicting polymer viscosity (Y) in a real-time manner and thus, determining when each batch should be terminated (end-point).

6.4.1 Data Set Description

The historical data set made available for the present study is composed of 47 batch runs. Each batch run contains the following variables described in section 6.1: Temp, Pres, R1W , R3W, WaterW, R1F, R3F, Water F and Valve. Additionally, four artificial variables were created to express the cumulative values of: coolant water and vapor flows used in the reactor jacket, gas purges through the vent valve and time consumption.

Preliminary data analysis indicated that reagent 2 flow and cumulative weight measurements were wrong, pointing to a complete sensor failure for all batches in the data set. These variables were thus excluded from the data set and plant personnel notified.

Additionally, flow and weight measurements relative to reagents 1, 3 and water, presented unjustified jumps for certain batches (probably due to the fact that these flow meters were also used to feed another reactor). These variables were thus filtered in order to eliminate false readings.

6.4.2 Batch Data Alignment

The structure of the data set under study does not allow for alignment according to traditional data trimming or indicator variable techniques. Data trimming is unsuitable due to the fact that measurements near the end of the batch, which are normally discarded when this methodology is used, are probably the most informative when end-point

prediction is intended. Additionally, indicator variables were not found for all phases of the batch run.

Thus, a simpler alignment method, proposed by Marjanovic *et al.* (2006), was applied. This technique consists of creating *pseudo batches* composed of batch-wise unfolded data, relative to, in this case, an hour before each viscosity sampling point. All data collected previous to this point is discarded. Consequently, in real-time, estimates of product quality (Y) will only be available during the condensation phase.

Figure 6.4 schematically represents the *pseudo batch* technique applied to the process under study. In this Figure, two batches are represented: in the first batch (1), the desired viscosity is achieved prior to the first sampling point; in the second batch (2), two samples had to be collected before this was verified.

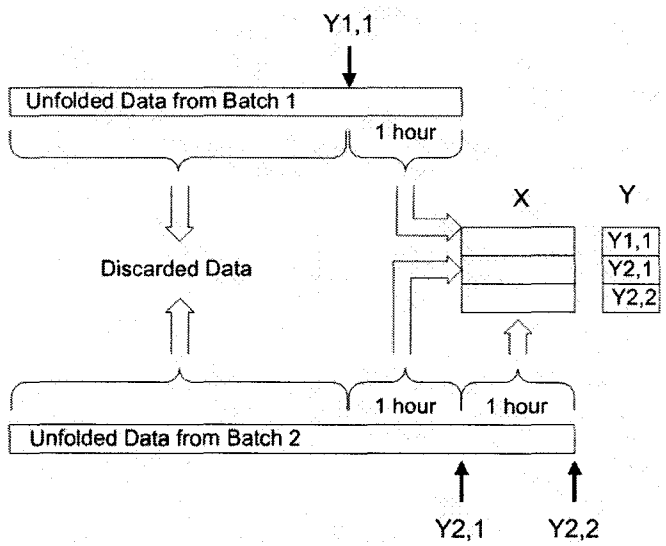

Figure 6.4 Representation of the *pseudo batch* technique, applied to the process under study.

A preliminary inspection of the viscosity data (Y) indicated discrepancies between the time in which a sample was claimed to be taken and that in which this actually occurred. From personnel experience, it is know that these differences can be as high as one hour for all samples, except for the first one. Thus, as a proof of concept, only

data relative to the first sample of each batch was kept. If the model proves to be useful in predicting Y, operators can be re-trained and new data collected.

6.4.3 MPLS

Using the pseudo batch approach

Following the generation of a data set composed of the *pseudo batches*, a standard MPLS model was built. Visual inspection of the resulting u1/t1 plot (Figure 6.5, left) points to a nonlinearity between the X and the Y data.

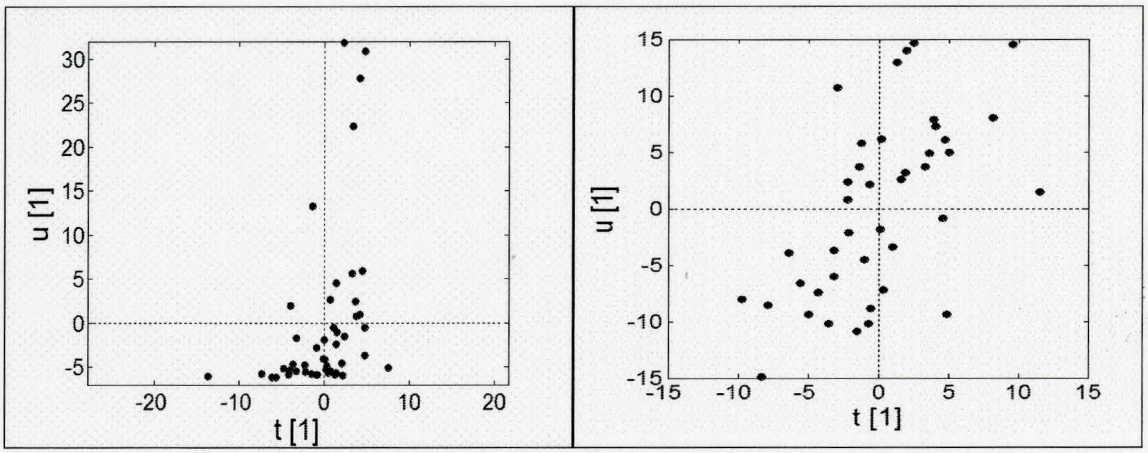

Figure 6.5 t1/u1 plots for MPLS models build using: Y = viscosity (left) and Y = log(viscosity) (right).

In order to compensate for this nonlinear behavior, a logarithmic transformation was applied to the response data, such that: Y = log (viscosity). Inspection of the u1/t1 plot (Figure 6.5, right), resulting from fitting a new MPLS model to the linearized data, set, showed that this objective was successfully achieved.

This new MPLS model is capable of explaining 54.3% (R^2Y) of the variations in the quality data and 21.8% (R^2X) in the process data using 2 PCs. However, it's capability of predicting the variations in Y is very poor (Q^2 = -21%). This indicates that

the process data used to build the model does not contain sufficient information to allow for a prediction of the polymer viscosity.

Using the X data throughout the whole batch run

Perhaps by discarding all data relative to the beginning of the batches, in order to create the *pseudo batches*, important information was also removed. With the intention of verifying if this occurred, a new model was built using the full process data set and the linearized viscosity data (Y). Since this is essentially an off-line study, the alignment technique applied consisted of identifying the different batch phases and using cumulative time to resample the data within them (as described in Chapter 3).

The resulting MPLS model is composed of 2 PCs and is summarized by: R^2Y = 72.8%, R^2X= 14.3% and Q^2 = -21%. Thus, this model also presents very poor predictive capabilities.

This indicates that, in fact, the process data set does not contain sufficient information to allow for the prediction of polymer viscosity and, consequentially, the end point of batches under production.

It is should be noted that the same Q^2 value was obtained by both models (using the *pseudo batches* and the full data matrix). This is due to the fact that the software used (simca-P+, version 10), truncates this value when it is smaller than -0.1. Thus, the Q^2 values presented should be understood as a qualitative indication of poor Y predictability.

6.5 Conclusions

In this chapter, a control methodology is initially proposed and implemented in order to specify a chemical admixture in terms of it's final application. Despite the success achieved by this methodology, it is based on manually sampling the process at specific time intervals, and thus represents a very time-consuming and dangerous

practice. Therefore, the use of MPLS models is explored, with the intent of predicting the polymer viscosity (Y) form the process data (X), in a real-time manner.

Due to the structure of the data set under study, the *pseudo batch* approach to data alignment (proposed by Marjanovic *et al.*, 2006), was initially applied. Additionally, a second MPLS model was built using data relative to the full batch runs.

None of these models were capable of predicting the variations in Y. This indicates that the process data (X), currently collected during the additive's production, does not contain sufficient information to predict this polymer's viscosity during a batch run. Consequentially, batch end-point times cannot be inferred. It is thus suggested that the process data set be enriched with measurements that indirectly reflect the consequences of an increase in polymer viscosity, such as the resulting electrical current in the agitator, in order for this objective to be achieved.

Chapter 7

Conclusions

Throughout this thesis multiway, multivariate, latent variable models and data preprocessing techniques were used in order to troubleshoot, monitor and predict data originating from industrial batch processes. In all, three different sets of data were studied and the main goal was to either reduce or infer product quality variability.

In the first data set analyzed (Chapter 3), MPCA and MB-MPLS methods were successfully used to troubleshoot an organic peroxide-producing batch unit in order to identify optimal process conditions with respect to quality. It was verified that most of the variability present in the Y data originated from a loose control in reagent addition.

Additionally, approaches to data laundering of the time-varying batch process variables, were proposed. To the best of the author's knowledge, this is the first time such issue has been discussed within the specific framework of batch analysis.

In Chapter 4, MPCA was successfully used to build a single, all-encompassing, on-line monitoring scheme for the heating phase of a multi-grade batch annealing process. In order to eliminate clustering due to the existence of multiple grades, grade specific mean centering was applied. The performance of this system was evaluated based on pre-established false alarm and fault detection metrics. Additionally, the MPCA model showed superior fault detection abilities when compared to the system that is currently in place at the annealing plant under study.

Issues relative to automated fault identification methods in batch process were also addressed within this work. It was verified that the distinction between certain fault

types, to which the annealing process is subjected to, is not clear. This is due to the fact that, at their initial points, these faults progress in a similar manner. Even so, a decision-tree approach was suggested to aid the operators or engineers in the fault identification process.

In Chapter 5, a data pre-alignment method was proposed and successfully used to synchronize batch data for on-line monitoring purposes. This technique is referred to as pre-alignment and it relies on a PLS model, fitted to variables obtained prior to the beginning of each batch cycle (initial conditions), to predict the duration of new batches. The predicted time information is then used to set the sampling rate of in-coming process data.

The performance of this technique relative to other, more traditional alignment methods, was also evaluated. For this purpose, an on-line MPCA monitoring scheme, built for the heating phase of a single-grade industrial batch annealing process, was used. The results of this comparative study indicated that data pre-alignment has a consistently superior performance relative to data trimming and an equivalent performance relative to crude linear interpolation and the use of an indicator variable.

Additionally, various methods for dealing with matrices containing different sized observations, in this case resulting from the existence of two possible steel stacking configurations, are proposed and evaluated. The method which presented best overall performance (highest predictive and predictor identification abilities) consisted of a single PLS model fitted to a data set in which measurements relative to inexistent coils were set as missing.

In Chapter 6, a control methodology is initially proposed and implemented in order to specify a superplasticizer in terms of it's final application. In spite of the success achieved by this methodology, it is based on manually sampling the process at specific time intervals and thus, a very time-consuming and risky practice. Therefore the use of MPLS models, with the intent of predicting the polymer viscosity (Y) form the process data (X) in a real-time manner, is explored.

Due to the structure of the data set under study, the *pseudo batch* approach to data alignment (proposed by Marjanovic *et al.*, 2006), was initially applied. Additionally, a second MPLS model was built using data relative to the full batch runs.

None of these models were capable of predicting the variations in Y. This indicates that the X matrices used to build these models, do not contain sufficient information to predict the polymer viscosity at a given time point. Consequentially, batch end-point times cannot be inferred from currently available process measurements. It is thus suggested that the process data set be enriched with measurements that indirectly reflect the consequences of an increase in polymer viscosity, such as the electrical current in the agitator.

References

Aitcin, P. "Concreto de Alto Desempenho", Editora Pini, LTDA, São Paulo, 2000

Berber, R., "Control of Batch Reactors: A Review", Methods of Model Based Process Control, ed. R. Berber. Kluwer Academic, Dordrech, The Netherlands, 459-494. 1995.

Bonvin, D., "Optimal Operation of Batch Reactors – A Personal View", Journal of Process Control, **8**, 355-368. 1998.

Buckley, A., Moses, A. J. and L. Trollope, "Study and Redesign of High Temperature Batch Annealing Furnace for Production of Grain Oriented Electrical Steel", Ironmaking and Steelmaking, **26**, (6), 477-482. 1999.

Champgne, M. and I. Ivanov, "Multigrade Modelling – Paperboard Quality Modeling", In: Proceedings of American Control Conference, Anchorage, AK. 2002.

Christie, O. H. J., "Data Laundering Leading to Meaningful Principal Component Classification Criterion and Map Attribute in Surface Chemistry", Aquatic Sciences, **57/3**, 242-254. 1995.

Christie, O. H. J., "Data Laundering By Target Rotation in Chemistry-Based Oil Exploration", Journal of Chemometrics, **10**, 453-461. 1996.

Clarke-Pringle, T. and J. F. MacGregor, "Optimization of Molecular-Weight Distribution Using Batch-to-Batch Adjustments", Industrial & Engineering Chemistry Research, **37**, 3660-3669. 1998.

Dong, D., McAvoy, T. J. and E. Zafiriou, "Batch-to-Batch Optimization Using Neural Network Models", Industrial & Engineering Chemistry Research, **35**, 2269-2276. 1996.

Eastment, H. T. and W. J. Krzanowski, "Cross-Validatory Choice of the Number of Components from a Principal Component Analysis", Technometrics, **24**, 1, 73-77. 1982.

Edgar, T. F., "Control of Unconventional Processes", Journal of Process Control, **6**, 99-110. 1996.

Eriksson, L., E. Johansson, N. Kettaneh, and S. Wold. Multi and Megavariate Data Analysis: Principles and Applications, Umea: UMETRICS AB , 1999.

Filippi-Bossy, C., J. Bordet, J. Villermaux, S. Marchal-Brassely and C. Georgakis, "Batch Reactor Optimization by Use of Tendency Models", Computers and Chemical Engineering, **13**, 35-51. 1989.

Flores-Cerrillo, J., "Quality Control of Batch Processes Using Latent Variable Methods", McMaster University Ph.D. Thesis, 2003.

Flores-Cerrillo, J. and J. F. MacGregor, "Control of Batch Product Quality by Trajectory Manipulation Using Latent Variable Models", Journal of Process Control, **14**, 539-553. 2004.

Friedrich M. and R. Perne, "Design and Control of Batch Reactors – An Industrial Viewpoint –", Computers chem. Engng., 19 Suppl, S357-S368. 1995.

Garcia-Munoz, S., T. Kourti and J. F. MacGregor, "Troubleshooting of an Industrial Batch Process Using Multivariate Methods", Industrial & Engineering Chemistry Research, **42**, 3592-3601. 2003.

Garcia-Munoz, S., "Batch Process Improvement using Latent Variable Methods", McMaster University Ph.D. Thesis, 2004.

Garcia, S. M., J. F. MacGregor and T. Kourti, "Product Transfer Between Sites using Joint-Y PLS", Chemometrics and Intelligent Laboratory Systems, **79**, 101-114. 2005.

Geladi, P. and B. R. Kowalski, "Partial Least-Squares Regression: A Tutorial", Analytica Chimical Acta 185, 1-17. 1986.

Jackson, J. E., A User's Guide to Principal Components, John Wiley, New York, 1991.

Juba, M. R. and J. W. Hamer, "Progress and challenges in batch process control", Chemical Process Control – CPC III, M. Morari and T.J. McAvoy (eds.), AIChE, CACHE Austin TX and Elsevier, Amsterdam, 139-183. 1986.

Kassidas, A., J. F. MacGregor and P. A. Taylor, "Synchronization of Batch Trajectories Using Dynamic Time Warping", Process Systems Engineering, **44**, (4), 864 - 875. 1998.

Kiparissides, C. and S. L. Shah, "Self-tuning and Stable Adaptive Control of a Batch Polymerization Reactor", Automatica, **19**, 225-235. 1983.

Kosanovich, K. A., Piovoso, M. J. and K. S. Dahl, "Multi-way PCA Applied to an Industrial Batch Process", In: Proceedings of American Control Conference, Baltimore, USA. 1994.

Kourti, T. and J. F. MacGregor, "Multivariate SPC Methods for Process and Product Monitoring", Journal of Quality Technology, **28**, (4), 409-428. 1996.

Kourti, T., "Multivariate Dynamic Data Modeling for Analysis and Statistical Process Control of Batch Processes, Start-Ups and Grade Transitions", Journal of Chemometrics, **17**, 93-109. 2003.

Kourti, T., P. Nomikos, and J. F. MacGregor, "Analysis, Monitoring and Fault Diagnosis of Batch Processes Using Multiblock and Multiway PLS", J. Proc. Cont., **5**, (4), 277-284. 1995.

Kozub, D. J. and J. F. MacGregor, "Feedback Control of Polymer Quality in Semi-Batch Copolymerization Reactors", Chemical Engineering Science, **47**, (4), 929-942. 1992.

Krzanowski, W. J., "Cross-Validatory Choice in Principal Component Analysis; Some Sampling Results", Journal of Statistical and Computational Simulation, 18, 299-314. 1983.

Lakshminarayanan,S., Gudi R. D., Shah, S. L., Nandakumar, K., "Monitoring Batch Processes using Multivariate Statistical Tools: Extensions and Practical Issues", IFAC Triennial World Cong., San Francisco. 1996.

Lee, K. S. and J. H. Lee, "Iterative Learning Control-based Batch Process Control Technique for Integrated Control of End Product Properties and Transient Profiles of Process Variables", Journal of Process Control, **13**, 607-621. 2003.

Lee, K. S., Chin, I. and H. J. Lee, " Model Predictive Control Technique Combined with Iterative Learning for Batch Processes", AIChE Journal, 45, 2175-2187. 1999.

Lennox, B., H. G. Hiden, G. A. Montague, G. Kornfeld, and P. R. Goulding, "Application of Multivariate Statistical Process Control to Batch Operations", Computers and Chemical Engineering, **24**, 291-296. 2000.

MacGregor, J. F. and P. Nomikos, "Monitoring Batch Processes", NATO Advance Study Institute for Batch Processing Systems Engineering, Antalaya, Turkey, Springer-Verlag, Heidelberg, 1992.

MacGregor, J. F., C. Jaeckle, C. Kiparissides, "Monitoring and Diagnosis of Process Operating Performance by Multi-Block PLS Methods with an Application to Low Density Polyethylene Production", AICHE Journal, **40**, 826-838, 1994.

MacGregor, J. F., "Data-Based Methods for Process Analysis, Monitoring and Control", Proceedings of IFAC System Identification SYSID'2003, Rotterdam. 2003.

Marjanovic, O. , Lennox, B., Sandoz, D., Smith, K. and M. Crofts, "Real-time Monitoring of an Industrial Batch Process", Chemical Process Control – CPC VII, Alberta, Canada. 2006.

Miller, P., R. E. Swanson and C. E. Heckler, "Contribution Plots: The Missing Link in Multivariate Quality Control", Appl. Math. And Comp. Sci., **8**, 4, 775-792. 1998.

Monte, R., "Avaliacao de Metologias de Ensaio Destinadas a Verificacao da Eficiencia de Aditivos Superplastificantes em Pastas de Cimento Portland", Universidade de Sao Paulo M. Eng. Thesis, 2003.

Moon, S. and A. Hrymak, "Scheduling of the Batch Annealing Process – Deterministic Case", Computers and Chemical Engineering, **23**, 1193-1208. 1999.

Muteki, K., MacGregor, J. F., U. Toshihiro, "Estimation of Missing Data Using Latent Variable Methods with Auxiliary Information", Chemometrics and Intelligent Laboratory Systems, **78**, 41-50. 2005.

Nelson, P. R. C., Taylor, P. A. and J. F. MacGregor, "Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations", Chemometrics and Intelligent Laboratory Systems, **35**, 1, 45-65. 1996.

Neogi, D. and C. E. Schlags, "Application of Multivariate Statistical Techniques for Monitoring Emulsion Batch Processes", Proceedings of the American Control Conference, USA, 1177-1181. 1997.

Nomikos, P. and J. F. MacGregor, "Monitoring Batch Processes Using Multiway Principal Components", AICHE Journal, **40**, (8), 1361-1375. 1994.

Nomikos, P., "Statistical Process Control of Batch Processes", McMaster University Ph.D.Thesis, 1995.

Nomikos, P. and J. F. MacGregor, "Multi-Way Partial Least Squares in Monitoring Batch Processes", Chemometrics and Intelligent Laboratory Systems, 30, 97-108. 1995a.

Nomikos, P. and J. F. MacGregor, "Multivariate SPC Charts for Monitoring Batch Processes", Technometrics, **37**, (1), 41-58. 1995b.

Perrin, A. R., Guthrie, R. I. L. and B. C. Stonehill, "Process Technology of Batch Annealing", Iron and Steelmaker, 15, (10), 27-33. 1988.

Qin, J. S., "Statistical Process Monitoring: Basics and Beyond", Journal of Chemometrics, **17**, 480-502. 2003.

Ramachandran, V.S., "Concrete Admixtures Handbook – Properties, Science and Technology", Ed. Noyes Publications, Park Ridge, New Jersey, USA, 2nd edition, 1997.

Sahay, S. S. and A. M. Kumar, "Applications of Integrated Batch Annealing Furnace Simulator", Materials and Manufacturing Processes, **17**, (4), 439-453. 2002.

Undey C. and A. Cinar, "Statistical Monitoring of Multistage, Multiphase Batch Processes", IEEE Control Systems Magazine, October edition, 40-52. 2002.

Westerhuis, J. A., Kourti, T. and J. F. MacGregor, "Analysis of Multiblock and Hierarchical PCA and PLS Models", Journal of Chemometrics, **12**, 301-321. 1998.

Westerhuis, J. A., T. Kourti and J. F. MacGregor, "Comparing Alternative Approaches for Multivariate Statistica Analysis of Batch Process Data", Journal of Chemometrics, **13**, 397-413. 1999.

Westerhuis, J. A., Kassidas, A., Kourti, T., Taylor, P. A., and J. F. MacGregor, "On-line Synchronization of the Trajectories of Process Variables for Monitoring Batch Processes with Varying Duration", submitted to AICHE Journal. 2003.

Wiel, S. A. V., Tucker, W. T., Faltin, F. W. and N. Doganaksoy, "Algorithmic Statistical Process Control: Concepts and an Application", Technometrics, **34**, 286-297. 1992.

Wold, S., "Cross Validatory Estimation of the Number of Components in Factor and Principal Components Models", Technometrics, **20**, 397-406. 1978.

Wold, S., N. Kettaneh, H. Friden, and A. Holmberg, "Modelling and Diagnostics of Batch Process and Analogous Kinetic Experiments", Chemometrics and Intelligent Laboratory Systems, **44**, 331-340. 1998.

Wold, S., P. Geladi, K. Esbensen, and J. Ohman, "Multi-Way Principal Components and PLS-Analysis", Journal of Chemometrics, **1**, 41-56. 1987.

Yabuki, Y., T. Nagasawa, and J. F. MacGregor, "An Industrial Experience With Product Quality Control in Semi-Batch Processes", Computers and Chemical Engineering, **24**, 585-590. 2000.

Yoon, S. and J. F. MacGregor, "Fault Diagnosis with Multivariate Statistical Models Part I: Using Steady State Fault Signatures", Journal of Process Control, **11**, 387-400. 2001.

Yoon, S., "Using External Information for Statistical Process Control", McMaster University Ph.D.Thesis, 2001.

Zavitsanou, A., "Robust Monitoring Models- Data Preprocessing Prior to Modeling", McMaster University M.A.Sc Thesis, 2002.