Pergamon

Computers chem. Engng Vol. 20, Suppl., pp. S745–S750, 1996

S0098-1354(96)00132-9
0098-1354/96 $15.00+0.00

EXPERIENCES WITH INDUSTRIAL APPLICATIONS OF PROJECTION METHODS FOR MULTIVARIATE STATISTICAL PROCESS CONTROL

THEODORA KOURTI, JENNIFER LEE AND JOHN F. MACGREGOR

McMaster Advanced Control Consortium, Chemical Engineering Department
McMaster University, Hamilton, ON, Canada L8S 4L7

Abstract - With process computers routinely collecting measurements on large numbers of process variables, multivariate statistical methods for the analysis, monitoring and diagnosis of process operating performance have received increasing attention. Recent approaches to multivariate statistical process control, which utilize not only the product quality data (as traditional approaches have done) but also the available process data, are based on multivariate projection methods (Principal Component Analysis, PCA, and Partial Least Squares, PLS). These methods have been rapidly accepted and utilized by industry. This paper gives a brief overview of these methods and illustrates their use for process monitoring and fault diagnosis with applications to a wide range of industrial batch and continuous processes. Emphasis is placed on the practical issues that arise when dealing with process data. Several of these issues are discussed and solutions are suggested for a successful outcome of the application of these methods in an industrial setting.

INTRODUCTION

Statistical Process Control (SPC) concepts have become very important in chemical and manufacturing industries. Their objective is to monitor the performance of a process over time in order to detect any special events that may occur. By finding assignable causes for them, improvements in the process and in the product quality can be achieved by eliminating the causes or, improving the process or, its operating procedures. Traditional SPC procedures, based on charting only a small number of final product quality variables, are totally inadequate for most modern process industries. They ignore the fact that with computers hooked up to nearly every industrial process, massive amounts of data are being collected continually on perhaps hundreds of process variables in continuous or batch processes. All such data should be used to extract information in any effective scheme for monitoring and diagnosing operating performance. However, all the process variables are not independent of one another. Only a few underlying events are driving a process at any time, and all these measurements are simply different reflections of these same underlying events. Multivariate statistical projection methods like Principal Components Analysis (PCA) and Partial Least Squares (PLS) are capable of utilizing massive amounts of data and compress the information in this data down into low dimensional latent variable spaces in which monitoring of the process and interpreting the results are much easier.

Kresta et al. (1991), MacGregor et al. (1991) and Wise et al. (1989, 1991) laid out the basic methodology of utilizing PCA and PLS for monitoring continuous processes. Extensions of these ideas to handle very large processes via multiblock PCA and PLS were made by MacGregor et al. (1994). Nomikos and MacGregor (1994, 1995b) used multi-way projection methods to develop methods for treating time varying batch processes. Kourti et al. (1995) combined multiblock and multi-way PLS to handle multistage batch processes. Once an unusual situation or fault has been detected it is important to diagnose an assignable cause for it. Miller et al. (1993) and MacGregor et al. (1994) proposed the idea of contribution plots for this purpose. Once a fault is detected the underlying PCA or PLS model is interrogated to calculate the contribution of each measured variable to the observed shift in the latent variables and prediction error. Recent reviews of developments in this area has been given by MacGregor and Kourti (1995), and Kourti and MacGregor (1995a,b).

The major advantages of these multivariate projection methods are their ability to handle large numbers of highly correlated variables, measurement errors, and missing data. Of particular importance is their ability to reduce the dimensionality of the monitoring space by projecting the information in the data down into low-dimensional spaces defined by a few latent variables. The processes are then monitored in these reduced spaces by using one-, two-, or three-dimensional control charts that retain all the simplicity of presentation and interpretation of conventional single variable SPC charts. However, by utilizing the information contained in all the measured variables simultaneously, they are much more powerful for detecting deviations from normal operating conditions.

The diagnostic procedures utilize the fact that underlying the multivariate monitoring scheme is a PCA or PLS projection model that characterizes the relationships among the process variables under normal operating conditions. Once a deviation is detected by the monitoring chart, the variables which make major contributions to this deviation are easily isolated using the underlying projection model. Once the major contributing variables are known, the diagnosis problem is much easier.

An important aspect of this multivariable statistical approach has been its rapid acceptance and utilization by industry. The following applications to continuous industrial processes have been reported. Wise et al. (1991) used PCA and PLS methods to analyze and monitor an industrial ceramic melter, Piovoso et al. (1992) used them to monitor an industrial polymerization process, Slama (1991) analyzed the operating behavior of the fluid bed catalytic cracking and fractionation section of a refinery, Miller et al. (1993) reported their use in diagnosing problems in photographic paper sensitization, Dayal et al. (1994) analyzed operating problems in an industrial pulp digester, and Hodouin et al. (1993) applied them to analyze the operating behavior of mineral crushing, grinding, and flotation circuits. Unpublished applications have been performed in the following industries: steel, tire and rubber, pharmaceutical, and chemicals.

Numerous applications of the multiblock and multi-way methods have also been reported on industrial batch processes. In particular, they have been widely applied to analyzing, monitoring and diagnosing various types of industrial batch and semi-batch polymerization processes: Nomikos and MacGregor (1995b), Piovoso et al. (1994), MacGregor and Kourti (1995), and Kourti et al. (1995). Unpublished applications include batch catalyst manufacturing, the annealing of steel, and the manufacture of microelectronics.

The objective of this paper is to illustrate the application of multivariate projection methods for process monitoring and fault detection and diagnosis on some new industrial applications to continuous and semi-batch processes. A variety of other applications will be used during the conference presentation to illustrate various concepts. Emphasis will be placed on the practical issues that arise with process data, and on ways of dealing with them in order to successfully apply these methods in an industrial setting.

OVERVIEW OF MONITORING BASED ON PROJECTION METHODS

Consider the situation where one has available a set of I multivariate observations on J process variables **X** (I x J) and L response (quality or productivity) variables **Y** (I x L). PCA can be used to decompose the **X** or **Y** into a set of A rank 1 matrices.

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^T + \mathbf{E}$$

where the loading vectors $\mathbf{p}_a$ define the reduced dimension space (A) with respect to the original coordinates and the score vectors $\mathbf{t}_a$ give the projection of the I observation vectors onto the A- dimensional reduced space. The vectors $\mathbf{t}_a$ are orthogonal to one another and are ordered with respect to variance. In PCA the vectors $\mathbf{p}_a$, $\mathbf{t}_a$ are the a-th eigenvectors of the matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ respectively. The number of significant dimensions (A) is usually selected via cross-validation, such that the residual matrix **E** contains no significant predictable component. The squared prediction error for the i-th observation vector is given by $SPE_i = \sum_{j=1}^{J} e_{ij}$.

In PLS the score vectors $\mathbf{t}_a$ and the associated loading vectors $\mathbf{w}_a$ are defined by both the **X** and **Y** matrices and represent a decomposition of the covariance matrix $(\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X})$. In this way the latent vectors $\mathbf{t}_a = \mathbf{X}\,\mathbf{w}_a$ represent the high variance directions in the **X** space that are most correlated with the important variables of interest in **Y**.

These methods can be utilized to analyze historical databases in order to better understand process behavior. In this case a PCA or PLS model is built from as wide a range of past operations as possible. By examining the behavior of the process in the score space of the dominant latent variables, regions of stable operation, sudden changes, slow process drifts, etc. can be readily observed, and often an interpretation for the changes can be found by examining the loading vectors or contribution plots. To employ these methods for monitoring processes, a PCA or PLS model must be built based on a different data set; one in which the process has been operating in an acceptable manner, and in which only good quality product has been obtained. In effect, one wants a model for the behavior of the process when it is operating well. The future operating behavior of the process is then monitored by comparing it against this model of good behavior. This involves plotting the scores (t_a; a = 1, 2, ..., A) and the SPE of new observations, and testing whether or not they are consistent with past good behavior. This is accomplished by establishing control limits for the monitoring charts using the statistical properties of the historical reference distribution of in-control data used to build the model. The monitoring is based on the concept that future observations on the process should continue to behave in a similar manner to the past data when the process is operating well.

For continuous processes, the **X** matrix consists of observations on J variables collected at successive time intervals. If the sampling interval is such that process dynamics are important then time lagged values of the process variables and/or output variables are included in the **X** matrix. Fundamental process model knowledge often is incorporated through the use of meaningful variable transformations and by including in the **X** matrix variables that are calculated from fundamental mass and energy balances, etc.

Batch processes are discontinuous in that they are of finite duration, and the variables vary about certain desired trajectories during the course of the batch. In a typical batch run j = 1, 2, ..., J variables (e.g.. temperatures, flows, pressure, etc.) are measured over k = 1, 2, ..., K time periods. Similar data exists on many different batches i = 1, 2, ..., I. These process data can be summarized in a three dimensional (I x J x K) array **X**. Also data may be available on L variables describing the final product quality taken at the end of each batch. These are summarized in the (I x L) matrix **Y**. Furthermore, for each batch information on initial conditions such as the analysis of feedstock properties, preprocessing conditions such as hold-up times, initial temperatures, and discrete operating conditions may be present. This information is summarized in an (I x M) matrix **Z**.

To analyze such data structures one can use multiblock, multi-way PLS. Multi-way methods allow one to decompose 3-dimensional data arrays into products of vectors and two-dimensional matrices. For the analysis and monitoring of batch processes this is most conveniently done as

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a \otimes \mathbf{P}_a + \mathbf{E}$$

where $\mathbf{t}_a$ is an (I x 1) score vector and $\mathbf{P}_a$ is a (K x J) loading matrix (Nomikos and MacGregor, 1994, 1995a,b). Prior to such a decomposition the original $\mathbf{X}$ array is first mean corrected by subtracting the sample means of each variable at each time and scaling the data (often to unit variance). In this way multi-way methods are looking at the variation of each variable about its mean trajectory. The additional preprocessing data matrix $\mathbf{Z}$ can be incorporated into the batch analysis using multiblock PLS (Kourti et al., 1995).

INDUSTRIAL APPLICATIONS

Two industrial applications are considered. A continuous process and a semi- batch polymerization process. These examples will serve to illustrate the methods, and the ideas of utilizing historical data to analyze the process and then utilize past good operation to set control limits for process monitoring. Several points critical for the successful application of the methods in an industrial setting will be discussed.

The first example is a continuous recovery process. The feed stream consisting of three major components A, B and C is passed through a series of separators. There are three products from the process, the most valuable being Product #1, high purity A. The principal operating objectives for the plant are to maintain the concentration of A in product #1 at a specified minimum level while achieving a certain minimum recovery of A to this stream. There had been instances of either low recovery or low purity in the plant. The company supplied historical data with the following objectives: i) to uncover the relations of purity and recovery to process variables and ii) to suggest a monitoring scheme for the plant. Daily averages on 447 process and product variables were provided for a period of 498 days. The data set had several missing data. During that period there were 7 days where the daily average values of concentration of A in product #1 were lower than the specified minimum value.

Multivariate projection methods were utilized for the analysis. PLS models were built between the process variables and purity and recovery. All the historical data were utilized for the initial model. The set of 442 process variables was projected to 7 principal components that could explain 93 % in purity variation and 93 % in recovery. Figure 1 shows the true values (solid line) of concentration of A in product #1, and the ones predicted (dotted line) by the PLS model (when there is no solid line true values were missing). The projection of the all the process data on the first two principal components is shown in Figure 2. Notice that all the points with concentration in A below the specified minimum of 99.5 % (marked as full circles) fall out of the acceptable operating region. Figures 3 and 4 show the contribution plots for two cases of low purity (points 479 and 480 in Figure 2). Notice that in both cases, out of 442 variables, the same group of variables appear to be related to the low purity product.

Multiblock analysis of the process determines the units with process variables related to a low purity product, and points to operating variables that might be responsible for that. Results from this analysis will be presented in the conference, with discussions on how to choose the set of data that could be used for the monitoring model. It is worth mentioning here that company engineers had spent considerable amount of time to determine what process conditions had led to low purity and /or recovery. The multivariate analysis pointed to those same variables, and in a very small fraction of the time spent by the engineers.

In the second example data were collected from 44 batches from an industrial semi-batch polymerization process. For each batch, the trajectories of 35 process variables measured during the run at 420 time intervals were provided, together with measurements of 6 quality variables obtained at the end of the run. Based on the value of the *measured* quality variables, all the batches were characterized by the company as normal (i.e., good). The product of this semi-batch process is subsequently used in several applications. Its performance in these applications however, may not be "observable" by the quality variables measured by the producer. Therefore, even if a batch is deemed good based on the measured quality variables, the projection methods may detect anomalies during the batch run (i.e., variations in the process variables) that affect product performance properties that are not otherwise observable.

This problem, where the measured quality variables are not enough to describe the product performance, is very common in industry. For example in the nylon production, the relative viscosity is the product quality variable usually monitored by the producer. However, other variables like the amine end groups affect the performance of the product in dyeing. By monitoring the relative viscosity only, one cannot access how the product will perform in dyeing. However by monitoring the process variables we were able to detect problems that were consistently affecting the dyeability of the product.

Although 35 variables were provided, only 11 variables were included in the model after discussion with the plant engineers. Variable selection is an important issue and may be an evolving process (try several models by pruning few variables each time and check the resulting model); careful variable selection will enhance the robustness of the final model that will be used for monitoring.

Another issue in the application of these methods to batch processes is a variable batch length. It is discussed by Nomikos and MacGregor (1995b) that in the three dimensional array (batch x variable x time), time can be replaced by another meaningful variable, as for example % conversion, or % of a component fed to the reactor in a semi-batch process. For this case the batches had a variable duration. However the total amount of ingredient "A" fed to the reactor was kept constant; therefore the progress of the batch, for this analysis, was not monitored against time but against the % amount of ingredient "A" fed to the reactor. One interval would correspond to 1% of the mass of "A" fed to the reactor.

Multi-way PLS was performed on all 44 batches to analyze this data base. The process variable measurements collected during the polymerization were arranged in a three-way array X (44x11x100), and the final product quality data were arranged in the matrix Y (44 x 6). A fast analysis indicated that one batch (#28) had an unusual behavior (large residuals compare to the rest of the batches). To establish a process monitoring scheme, this unusual batch was omitted from the data set and a new multi-way PLS model was attempted using the remaining "good" batches. When attempting building a PLS model, no component was cross-validated. This result means that quality was so consistent, that the only variation in the quality properties was due to random process noise and measurement reproducibility. In that case the best approach is to built a PCA model to define normal operating conditions for the process. A MPCA model with four components was able to explain 55.0 % of the process variation. Based on the model residuals from this reference distribution of good batches, control limits were established for the scores and the SPE (Nomikos and MacGregor, 1995). New batches are then monitored by computing the scores and the SPE at each time interval k and checking that they lie within their control limits.

Figure 5 shows some of the monitoring charts for batch # 28. The batch run appeared to be progressing well until about time period 74 when the SPE and t_1, t_4 exceeded their control limit. To help diagnose a reason for the observed deviations, the underlying MPCA model can be interrogated to compute the contribution of the individual variables to SPE, t_1 and t_4, at interval 74. These contributions are plotted in Figures 6a-6c. It appears that the "ingredient C flow rate " is the main contributor. The trajectories of this variable for all 43 batches in the original data set are shown with grey background in Figure 6d and for batch #28 as solid line. At interval 74, for batch #28, the value of this variable increases rapidly, compare to other batches, and remains large for some time. Investigating this behavior, it was discovered that the operator had typed the wrong value for the set point of the flow of ingredient C, and did not notice the mistake for sometime. Because the event happened near the end of the run no effect on the product quality was observed. However, had this anomaly occurred during the first 20 observations it would had very serious implications in the product quality. With on-line use of these methods, such mistakes would have been caught and alarmed instantaneously, avoiding disastrous effects on the product quality.

SUMMARY

Process monitoring, early fault detection and diagnosis, is important in chemical an manufacturing industries for safety, controlling product quality, minimizing waste of raw materials, etc. Large multivariate processes are difficult to monitor by traditional methods. The simplicity of the presentation and interpretation of the multivariate charts based on the projection methods discussed here, makes these charts attractive to the plant engineers and operators. The development and application of these methods to industrial processes is one of the most rapidly growing topics in the systems and control area. This paper has reviewed much of the published theoretical and industrial applications literature on this topic. Two industrial applications to a semi-batch polymerization process and a continuous process were presented to illustrate the power of these methods in analyzing, monitoring and diagnosing operational problems. Issues that need consideration for their successful application in industrial settings, are also discussed. These issues and others that will be presented during the conference come from our experience with applying these techniques to more than 30 industrial processes, both batch and continuous. These processes were from diverse industries such as the pharmaceutical and semiconductor industry to steel, polymer and petroleum industry. Several companies have started utilizing these methods on production scale operations and an update on their success will be discussed in the conference.

REFERENCES

Dayal, B.S., J.F. MacGregor, P.A. Taylor, S. Marcikic and R. Kidlaw, 1994, *Pulp and Paper Canada*, **95**, 26.

Hodouin, D., J.F. MacGregor, M. Hou and M. Franklin, 1993, *Canadian Institute of Metallurgy Bulletin*, **86**, 23.

Kourti, T., and J.F. MacGregor, 1995a, *Chemometrics and Intelligent Laboratory Systems*, **28**, 3.

Kourti, T., and J.F. MacGregor, 1995b, *Journal of Quality Technology*, accepted for publication.

Kourti, T., P. Nomikos and J.F. MacGregor, 1995, *Journal of Process Control*, **5**, 277.

Kresta, J., J.F. MacGregor and T.E. Marlin, 1991, *Canadian Journal of Chemical Engineering*, **69**, 35.

MacGregor, J.F., C. Jaeckle, C. Kiparissides and M. Koutoudi, 1994, *AIChE J.*, 40, pp. 826-838.

MacGregor, J.F., and T. Kourti, 1995, *Control Engineering Practice*, Vol. 3, No. 3, pp. 403-414.

Miller, P., R.E. Swanson and C.F. Heckler, 1993, 37th Annual Fall Conference, ASQC, Rochester, NY.

Nomikos, P., and J.F. MacGregor, 1994, *AIChE Journal*, Vol. 40, pp. 1361-1375.

Nomikos,P., and J.F. MacGregor, 1995a, to appear in *Chemometrics and Intelligent Laboratory Systems.*

Nomikos, P. and J.F. MacGregor, 1995b, *Technometrics*, Vol. 37, No. 1, pp. 41-59.

Piovoso, M.J., K.A. Kosanovich, K.S. Dahl, J.F. MacGregor and P. Nomikos, 1994, *American Control Conference*, Baltimore, Maryland, June 29-July 1, 1994.

Skagerberg, B., MacGregor, J.F., and Kiparissides, C., *Chemometrics and Intelligent Laboratory Systems,* **14**, 341-356 (1992).

Slama, C. (1991). *Analysis of Industrial FCCU data using PCA and PLS.* M.Eng. Thesis, McMaster University. Hamilton, Ont., Canada.

Wise, B.M., and N.L. Ricker, 1989, *AIChE Symposium Series*, Vol. 85, No. 267, pp. 19-23.

Wise, B.M., D.J. Veltkamp, N.L. Ricker, B.R. Kowalski, S.M. Barnes and V. Arakali, 1991, Application of Multivariate Statistical Process Control (MSPC) to the West Valley Slurry-Fed Ceramic Melter Process", Waste Management '91 Proceedings, Tucson, AZ, 1991.

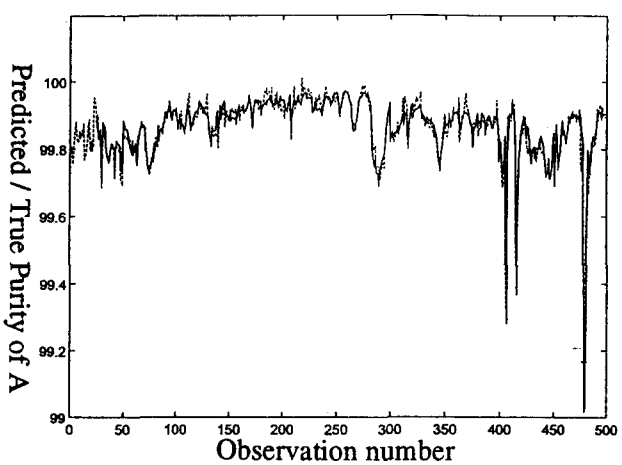

Figure 1. True (solid line) and PLS model predicted (dotted line) value for concentration of A in product #1.

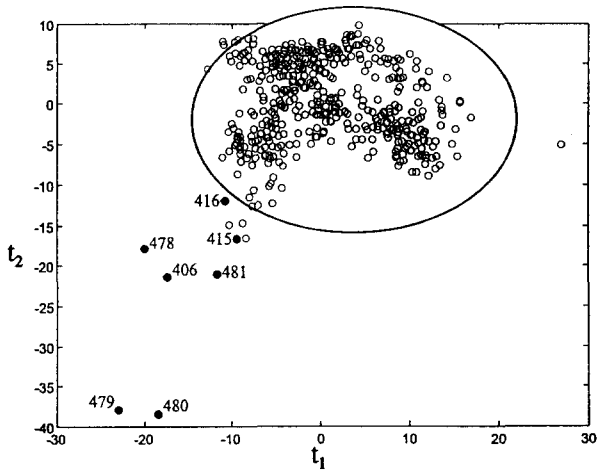

Figure 2. Projection of 498 days of operation on t_1-t_2 plane. Solid circles are for concentrations of A below the 99.5 % specification limit

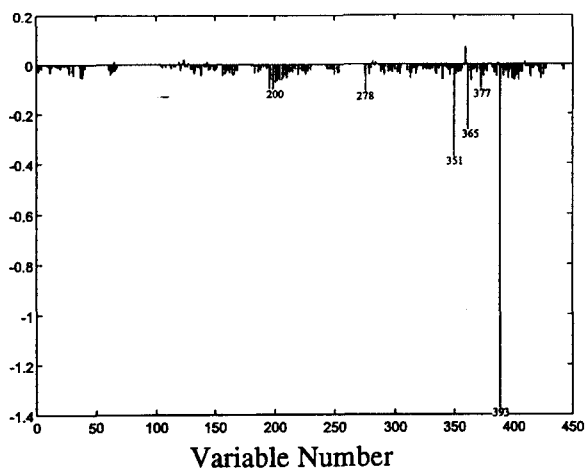

Figure 3. Process variable contributions to observation 479 with A concentration below 99.5 %. Variables 351, 365, 393 have a significant contribution.

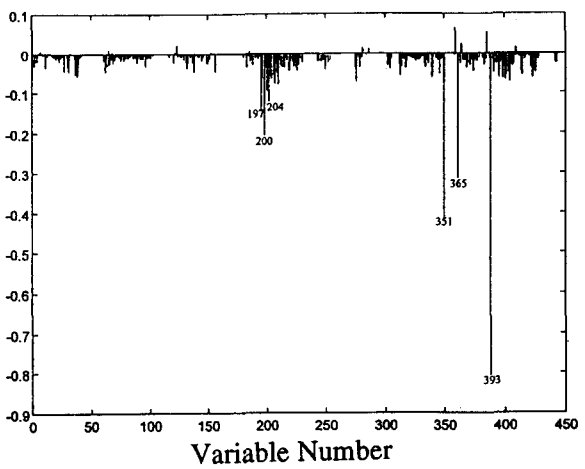

Figure 4. Process variable contributions to observation 480 with A concentration below 99.5 %. Variables 351, 365, 393 have a significant contribution.

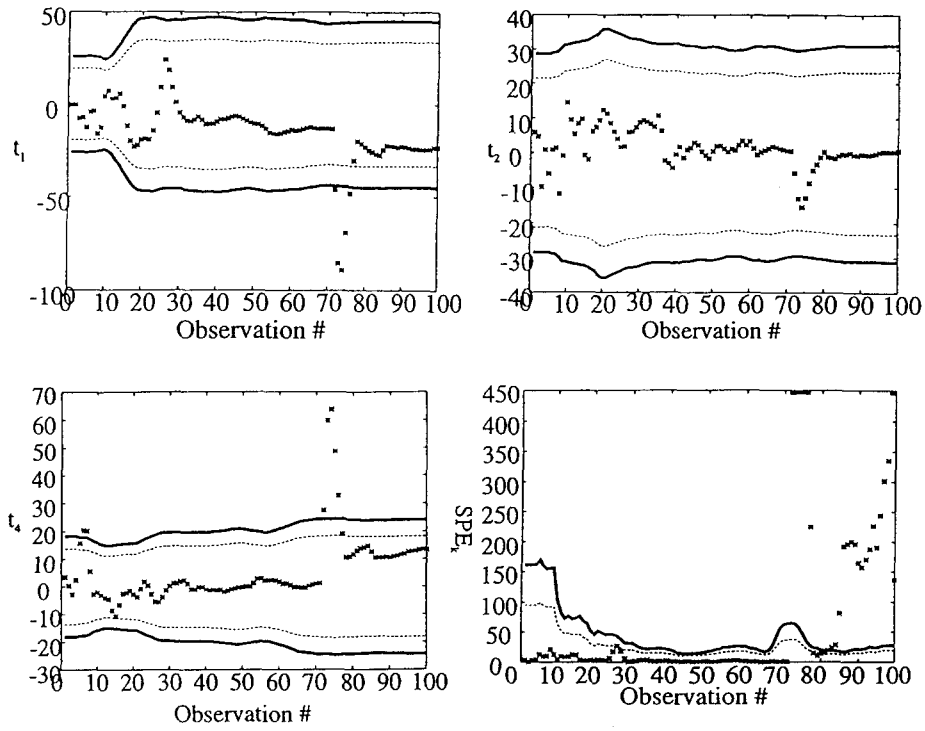

Figure 5. Control Charts for Batch #28.

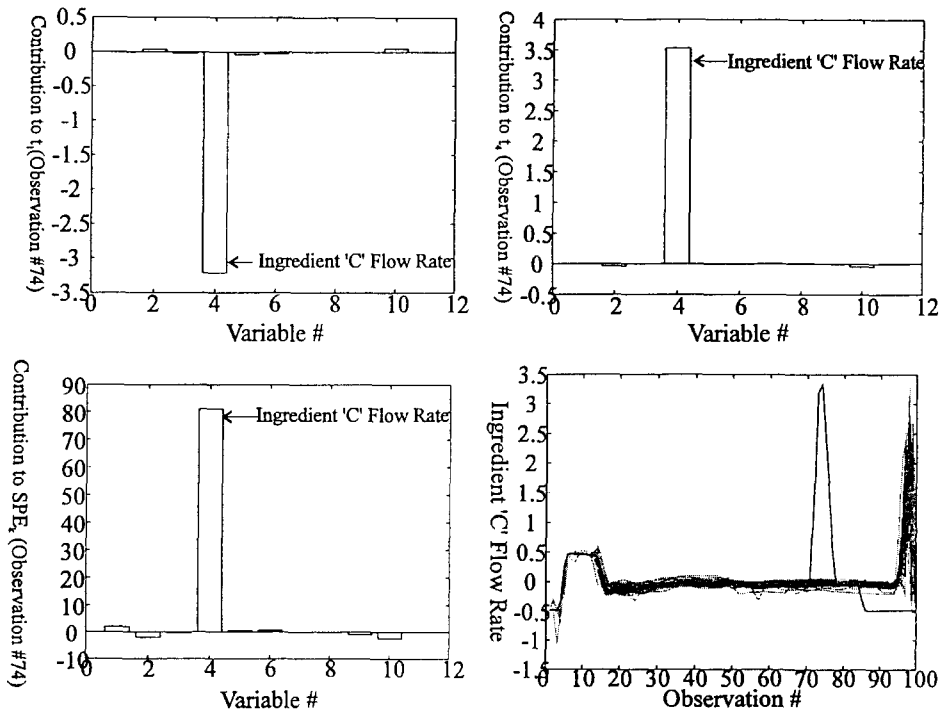

Figure 6. Fault Diagnosis for Batch #28. a-c) contribution plots to t_1, t_4, SPE. d) Trajectory Flow-rate of Ingredient C for Batch #28 (solid line) against the other 43 batches.