# Batch Process Analysis using MPCA/MPLS, A Comparison Study

Salvador Garcia
Dept. of Chemical Engineering, McMaster University, Hamilton, Ontario, L8S 4L7, Canada

***Abstract* – Batch processes can be monitored using MPCA, two different approaches have been proposed. Both methods are compared from a fundamental and practical point of view. The scheme proposed by P. Nomikos and J.F. MacGregor gives a fine and robust process monitoring tool, the approach proposed by Wold *et al* gives a coarse monitoring tool and can lead to an "insensitive" control chart that doesn't register all the process upsets and deviations from the desired performance. PLS can be successfully used for inferring properties of the batch if the PLS model has a fundamental background and meaning. The loadings of the calculated principal components can serve as a guide in building a business case when a MSPC solution is to be implemented.**

## I. INTRODUCTION

SPC has been used for several decades for process monitoring. Kresta, MacGregor and Marlin proposed the use of multivariate statistics for process monitoring (Kresta, MacGregor, Marlin, 1991). Multivariate SPC (MSPC) offers the possibility of analyzing the data coming from a DCS or from a real-time database and monitor closely the performance of the process. MSPC uses Principal Component Analysis (PCA) and Partial Least Squares (PLS) in order to monitor the process.

PCA and PLS have been used successfully in many continuous industrial processes. Multi-Way PCA and PLS (MPCA, MPLS) was proposed by S. Wold *et al* (Wold, Geladi, Esbensen & Ohman, 1987) as an approach for batch process analysis, later P. Nomikos and J.F. MacGregor (Nomikos and MacGregor, 1994,1995,1995b) proposed a formal methodology for batch processes MSPC, using also MPCA and MPLS.

Wold and collaborators (Wold, Kettaneh, Fridén, and Holmberg, 1998) proposed later their own batch process MSPC methodology, this method is used within SIMCA-P from Umetrics.

The aim of this work is to compare both methods (Wold and Kettaneh, from Nomikos and MacGregor) in the effectiveness, robustness and ease of use, for monitoring an industrial process.

The work is organized in the following way: First, both methods are presented and described; second, some fundamental issues are discussed; third, two industrial cases are presented; forth, PLS is discussed for inferential sensors in batch processes; fifth, we discuss how to build a business case, and last, some conclusions and future work.

## II. BATCH MSPC METHODOLOGIES

Batch processes are, by nature, a 3D matrix of data (Fig. 1.) MPCA needs to unfold this matrix in order to obtain a 2D matrix of data, and then performing PCA/PLC. For unfolding this matrix, first it is divided into K "slices"; each resulting slice is a matrix with I rows and J columns. These slices are going to be unfolded. The direction of this unfolding will decide the mean centering direction and therefore the complete interpretation of the resultant principal components. The direction of this unfolding is the main difference between the methods to be compared.

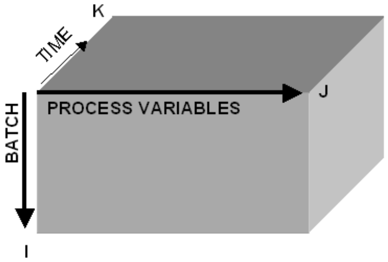


Fig.1. Batch data is a 3D matrix of data

### A. Nomikos and MacGregor approach.

The unfolding proposed by the authors of this methodology is to be done time wise, so the resultant 2D matrix has I rows and (K X J) columns (Fig.2.). Then the matrix is mean centered, this is, to calculate the mean for each column, and then subtract it from the variables. This mean centering will remove the mean trajectory of each variable, and the monitoring of future batches will be done around this mean trajectory.

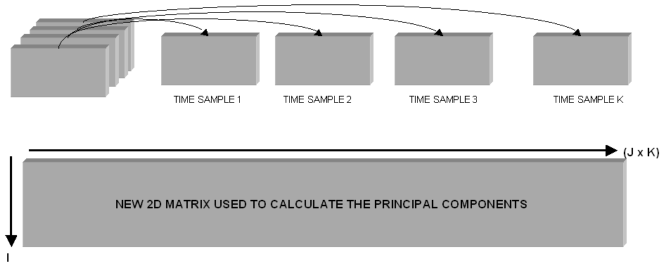


Fig.2. Unfolding done time wise

The main objective of any automation system driving a batch process is to maintain the batch performance around the desired trajectory, so it will give on-spec product. The mean centering done with this method is alike a monitoring scheme around the mean trajectory of its variables (Fig 3).

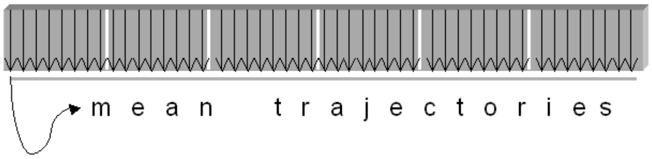


Fig.3. The mean centering is done around the mean trajectory.

Once the matrix is mean centered and scaled, PCA is performed, the results from PCA are the loading vectors, and the calculated scores for each batch (Fig 4.). The loading vectors will have a weight for each variable at each time, so, when used for monitoring of future batches two problems arise (Fig.5.):

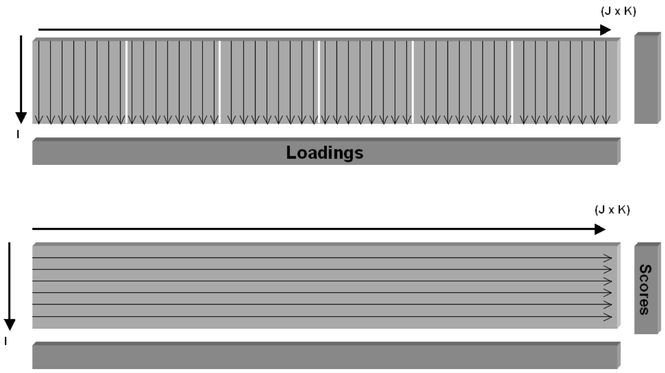


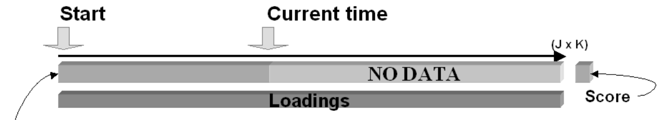
Fig. 4. Loadings and scores for the unfolded matrix

*1) Missing data during monitoring:* When a batch is been monitored, we know the values from the beginning and all the way to the current time, but there is a complete block missing in order to compute the value of the scores (and other statistics) at this current time. Nomikos and MacGregor suggest three different ways of dealing with this problem:

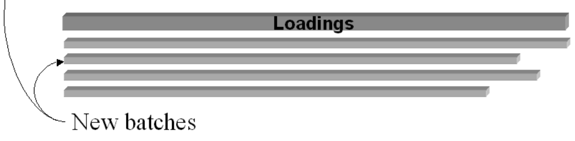
a) Fill the missing values with the current values.
b) Fill the missing values with zeros.
c) Use the ability that PCA has to handle missing data.

The implications of using one or another will be discussed later, the limits of the control charts will depend on the method used to deal with missing data.

*2) Batch lengths:* The calculated loadings can be used for monitoring, only if we assume that all the following batches will have the same number of samples; this almost never happens, so, we need to use some kind of alignment for the batches.



*i. Missing data during monitoring*



*ii. Batch unequal time durations*

Fig. 5. Problems encountered when monitoring new batches.

The most accepted idea is to use another variable, different from time, in order to mark the beginning and end of each batch, this variable has to grow monotonically during the batch, and has to have the same value at the end of all batches, one good example is conversion.

The last method for synchronizing batches was proposed by MacGregor (Kassidas, MacGregor and Taylor, 1998) in this work they use the philosophy behind dynamic time warping (DTW) in order to align batches with different duration. DTW is a method used in computer science for speech recognition; a computer has to be able to recognize a word, regardless of the speed used to pronounce it, DTW is used to reorder the digitalized sound and fit it into a known pattern.

The control charts come naturally since the values of the scores can be calculated as the batch advances. The limits for these charts are discussed in Nomikos and MacGregor, 1995.

### *B. Wold and Kettaneh method.*

In this case the proposed unfolding is done batch wise, so the resultant matrix has (I X K) rows, and J columns. The matrix is mean centered batch wise (Fig.6.), so the "great mean" is removed from each variable.

Doing so is like monitoring the batch around the mean of the variables during all the batch, this is rarely done in industry because a batch may keep its trajectory within the this limits, but give off-spec product, more issues regarding mean centering will be discussed later.

After mean centering and scaling, a new vector is added arbitrarily to mark the maturity of the batch; this vector contains the timestamp of each sample (Fig.7), this vector can be replace for another one if available (like conversion), this vector is to be considered as the Y vector, and a PLS model is done to model the process variables and the new Y vector; the validity and significance of doing so is also to be discussed later.

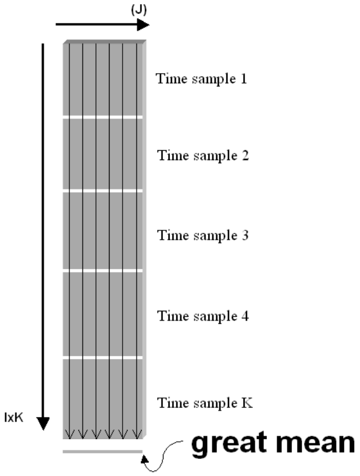


Fig. 6 Mean centering batch wise

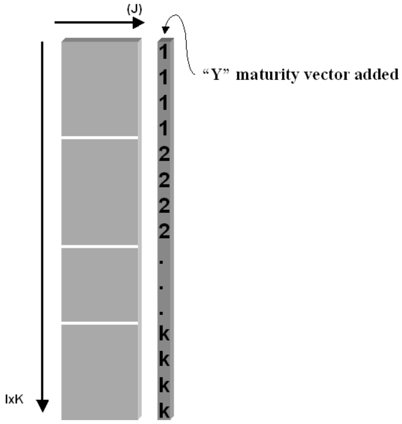


Fig. 7. Maturity vector added to the data.

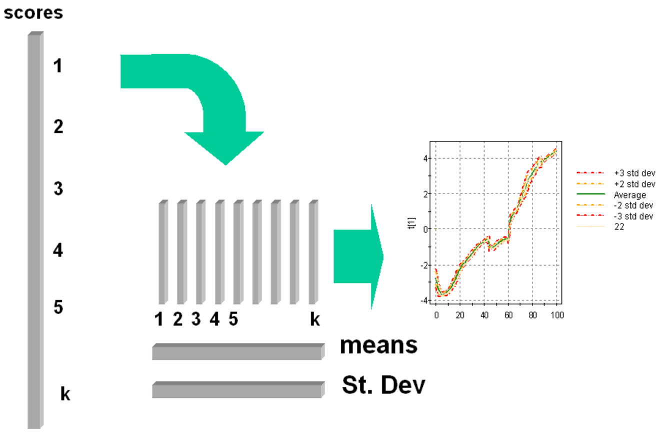


Fig.8. Scores are reordered time wise in order to obtain the control charts.

The resultant scores will show the difference between the time samples, and the loadings will reflect the correlation and importance of the process variables from an "overall" perspective, without taking in account the time dependency and correlation.

In order to obtain the control charts, the scores of the model are reordered time wise, since the monitoring has to be time dependant, and a control chart is built based in the time wise mean of the scores, and their standard deviation (Fig. 8). Doing this, the method tries to correct the fundamental mistake done when building the model.

## III. FUNDAMENTAL IMPLICATIONS

As mentioned before, there are some fundamental problems with doing the unfolding of data batch wise. The nature of the model, the quality of the monitoring, the weight given to different stages of the batch, and the interpretation of the final scores and loadings are some of them, and will be discussed in detail.

### *A. Nature of the model.*

The monitoring problems we will face when using the Nomikos-MacGregor approach are both related to the time dependency of the model. These problems are not present in the Wold-Kettaneh scheme because the resultant model doesn't include the time dependency of the correlation structure.

Batch processes are by nature a dynamic process, and the correlation among the variables depends on time, so, any model that claims to represent the correlation between variables for a batch process must have this time dimension included. The loadings of the resultant model from unfolding batch wise do not include any variable that could account for the time dimension; it doesn't make much sense to represent a time-dependant process with a non-time-dependant model.

### *B. Quality of monitoring*

A profitable batch process has gone thru a lot of refining in the design (that is why it is profitable) in order to design the trajectory that the batch is supposed to follow. Especially in the polymer area, a small change in the trajectory may lead to off-spec product, so it is important to perform a fine monitoring scheme on the batch trajectory.

When unfolding time wise, the mean centering and scaling leads to data that represents the deviation of the batch from the mean trajectory, and this is precisely what the loadings and scores will model, the deviation of each time sample from the desired mean trajectory (Fig. 9)

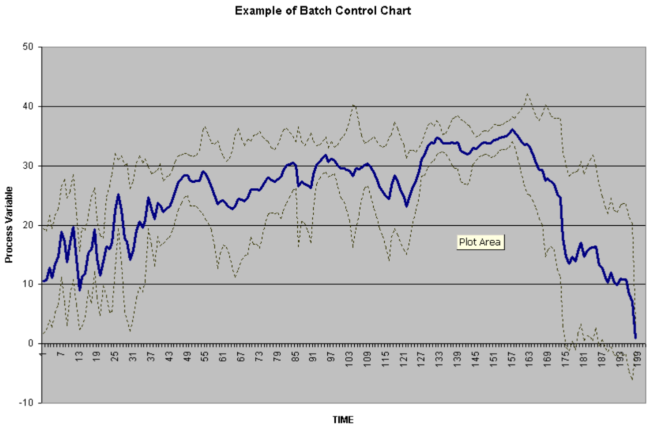


Fig. 9. Monitoring chart around the desired trajectory

If the unfolding is done batch wise, the mean centered and scaled data will represent the deviation of the variables from the great mean (this is the average value of each variable during all the batch) and this is what we will be modeling with the loadings (Fig. 10).

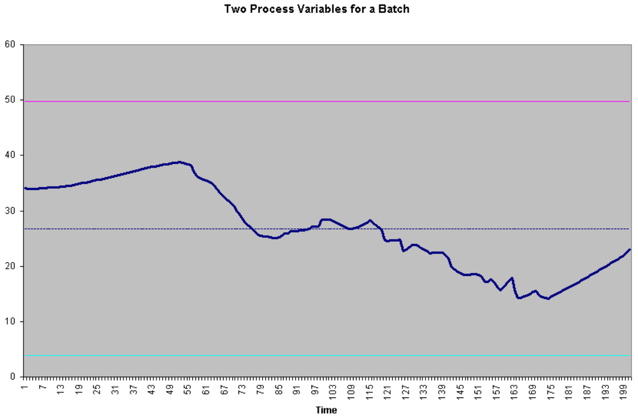


Fig. 10. Monitoring chart around the great mean

It is quite obvious that the best monitoring scheme will be the one obtained with time wise unfolding, and this will be seen with the industrial examples presented further in this work.

### *C. Weight given to different stages of the batch*

Batch processes often have several stages, (i.e. initiation, propagation, termination) or maybe the addition of a second catalyst to the reaction, or the presence of a strong exothermic period. These stages may have different impact in the final product quality, and most of the time they are separated by a period of randomness between variables, because of the absence of the driving force of the batch (Fig. 11).

When the unfolding is done batch wise, the mean calculated takes in account this "random walk" periods. Hence, it is giving the meaningful periods the same importance that the noisy ones. In the Wold-Kettaneh approach, when the scores are time-wise reordered to calculate the means, and standard deviations they try to recover from this, but the "damage" is already done because the loadings are biased by this noisy batch periods.

In the Nomikos-MacGregor approach, the unfolding is done time wise, and therefore we are giving freedom to the PCA or PLS method to find this "random walk" periods.

We will expect the main variation in the process to be because of variations in the meaningful stages, rather than the noise (i.e. an impurity in the raw materials will affect the kinetics) so, if we analyze the loadings and the remaining sum of squares per time, we will be able to distinguish this periods of the batch where the correlation is just noise, because it will not be captured as part of the latent events occurring in the batch (Fig. 12).

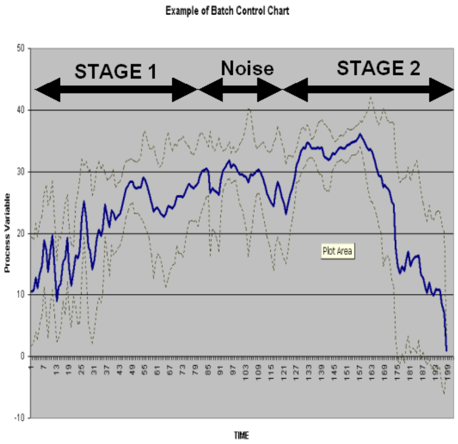


Fig. 11. Different stages of a batch

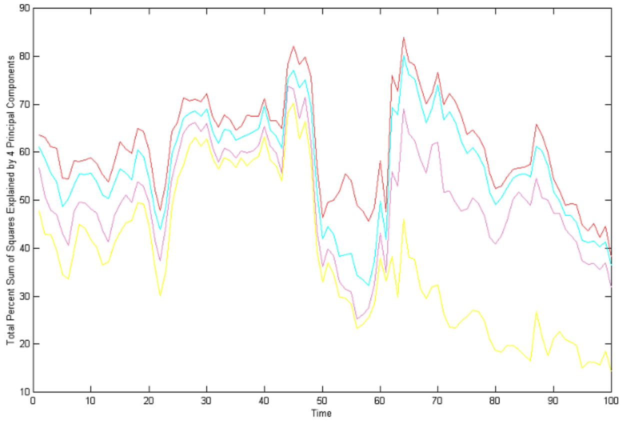


Fig.12 Captured sum of squares shows a noisy period that is not been captured by the model.

### *D. Interpretations of the loadings and the scores*

In the Nomikos-MacGregor method (Fig.4), the scores calculated for each batch can be seen, as a "measure" of how different are batches from each other, each batch has a single "measure" that can be used to qualify a batch as "good" or "bad" (Fig. 13).

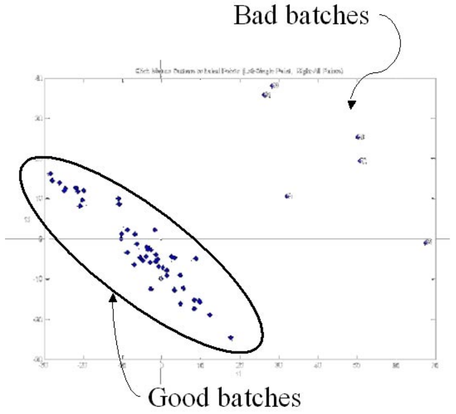


Fig. 13. Score plot clearly shows if a batch is normal or not.

The loadings calculated will reflect the correlation and importance between variables during time, and they will be easily interpreted in terms of the fundamental phenomena occurring in the batch, i.e. if original data includes temperatures and pressures, the loadings can be interpreted in terms of the energy balance around the system.

Because of the time dependency of the model, the loadings can be easily interpreted, and their interpretation will offer deep understanding of the batch itself.

This is a very important place where statistic and mechanistic models come together, the loadings of the principal components must reflect the fundamental phenomena occurring in the batch, and, in a PLS model, will represent the fundamental correlation between the X and the Y space.

In the Wold-Kettaneh methodology, each time sample will have its score in such way that there will not be any single value to determine if the batch is good or not, the need of analyzing trajectories is still there in order to see if a bath is "good" or "bad" (Fig. 14).

The interpretation of the loadings obtained with the Wold-Kettaneh method will give an overall picture of the correlation between the batch variables but will not tell anything about the time relevance of them. So, in the best of the cases, we will be able to say that "a pressure is correlated with a temperature" and most of the times we already know that.

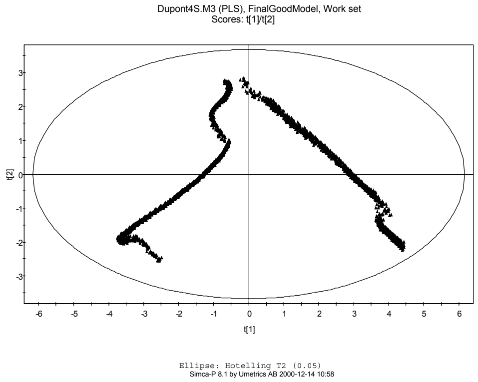


Fig. 14. Score plot obtained when the unfolding is batch wise, clustering is not visible

## IV. TWO INDUSTRIAL APPLICATIONS

Data was taken from two polymer plants, one from a nylon process and another from a PVC process. Sets were analyzed using the methods described earlier, the software used was SIMCA-P, this program has the method by Wold-Kettaneh, and BatchSPC has the Nomikos-MacGregor approach. Besides the efficiency and ease of use of each method, other issues are analyzed like a rigorous solution to the problem of missing data when monitoring, and the validity of an inferential sensor during the batch using PLS.

In each case we will walk thru the following path:

a) Make a PCA model with raw data
b) Look for outliers or clustering.
c) Remove outliers and make another PCA model
d) Analyze contributions for outliers or clusters.
c) Analyze remaining variance (or sum of squares) or other statistics (Hotelling T2).

### *A. Data from a nylon process.*

Data set consists of 55 batches, 10 variables were sampled 100 times for each batch.

The first two principal components are obtained; Fig. 13 is the score scatter from BatchSPC and shows a clear clustering of batches 50 to 55, while in the SIMCA-P graph (Fig.15) instead of a clustering, we see batches 50-55 lying outside the control limits for the DModX chart.

Removing batches 50 to 55 and redoing models, we will see no clearer clustering in the BatchSPC t1t2 plot (Fig. 16), but in the DModX chart (Fig. 17) we clearly see that batch 49 is very distant from the model, while in the SIMCA-P control chart, this is not easy to see (Fig.18).

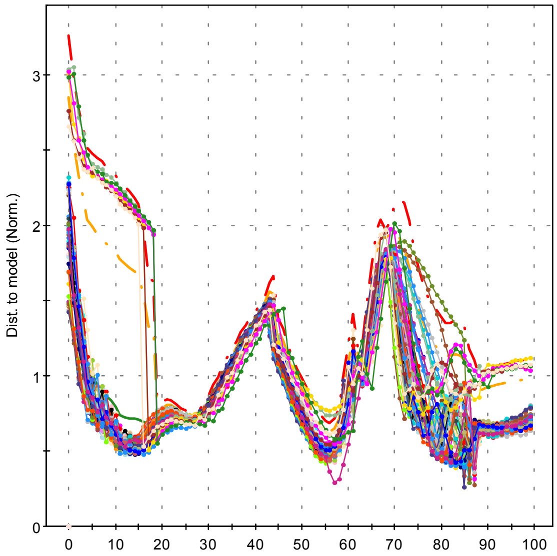


Fig. 15 Control chart from SIMCA-P showing batches 50 to 55 outside the control limits

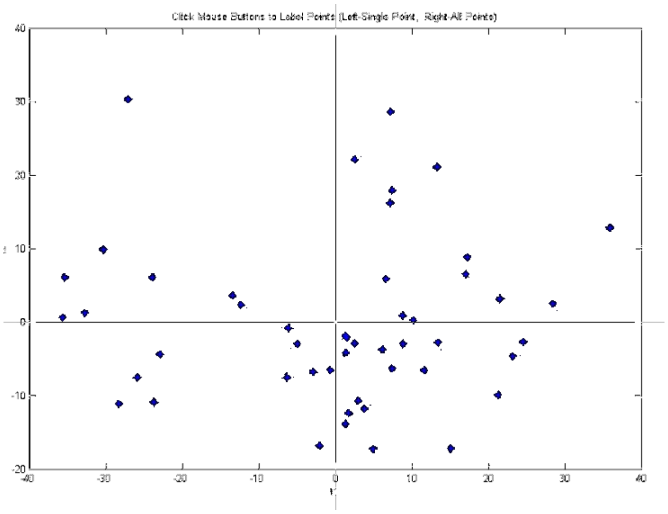


Fig. 16 BatchSPC t1t2 chart without batches 50 to 55

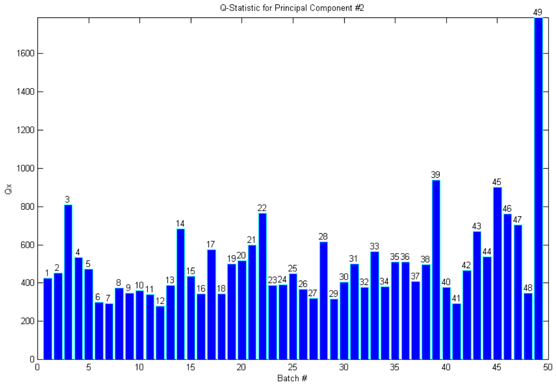


Fig. 17 Remaining sum of squares showing batch 49 to be different from all the others

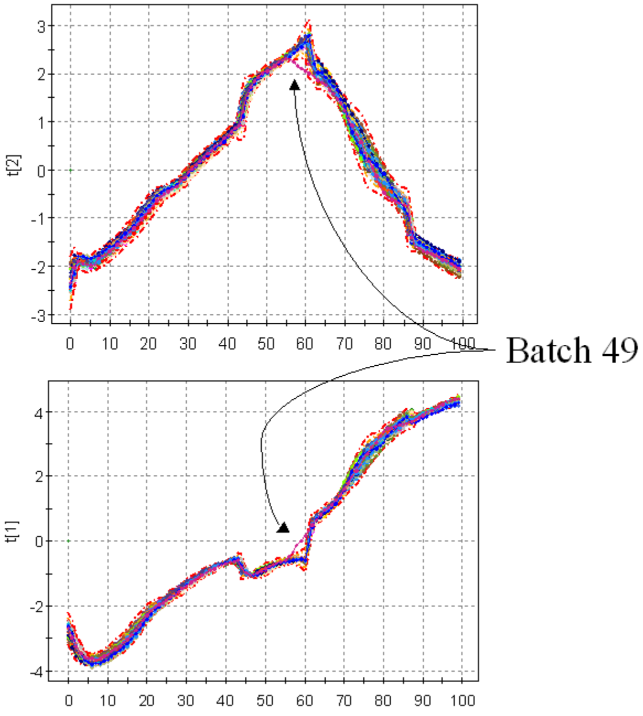


Fig. 18 Looking at the SIMCA-P control charts its not obvious that batch 49 is different

In fact, once all the know batches are removed (using previous knowledge from data owner) we still see some clustering in the BatchSCP t1t2 score (Fig. 19) and we also can distinguish batches 22 and 14 as different from both clusters, while in the SIMCA-P this clustering is not visible (Fig. 20). The fact of having one score per batch – in BatchSPC- gives the possibility of studying the contributions of specific variables, in order to find more about the differences of a certain batch from a cluster of batches, from another batch or from the latent variable model.

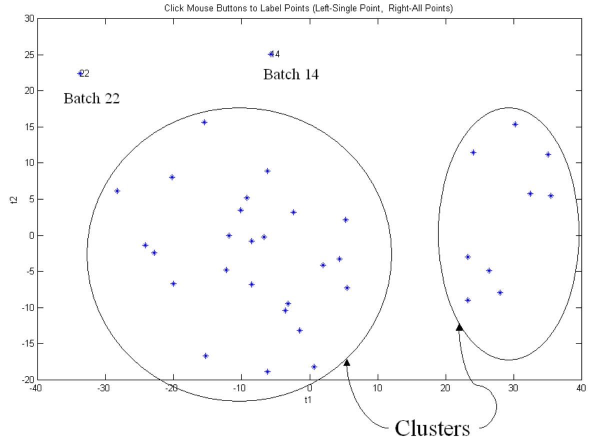


Fig. 19 BatchSPC t1 t2 score for good known batches

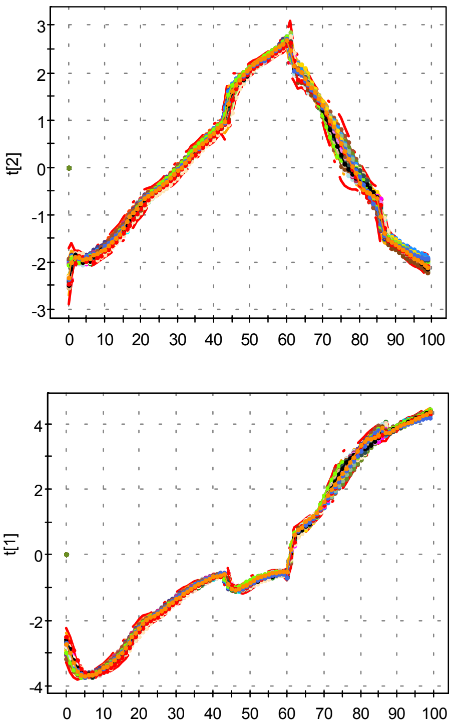


Fig. 20. SIMCA-P control charts only for good known batches.

In this point of the analysis, the Wold-Kettaneh method has lost analysis capability; with the Nomikos-MacGregor model we can go further still. Since this approach does consider time, the contributions we can calculate will show "when" a variable was important for the deviation, i.e. for batch 22 the variable # 1 was important only during the first phase or stage of the batch, and variable # 8 was important only for the second stage of the batch (Fig 21).

A PLS model tries to find the greatest variation dimension in the X space that explains the greatest one in the Y space, and it assumes a linear relationship between both latent spaces, a graph for the scores in the X and the Y spaces should show a linear relationship.

Wold-Kettaneh method does a PLS model between the X matrix and a "maturity" vector. Fig 22 plot has t1 vs. u1 showing clearly a non-linear relation. The predictions of time have multiple solutions and are far from a monotonic growing indicator of maturity.

A PLS model should have a fundamental reason to predict the Y space, and if we look closer to the PLS model built in the Wold-Kettaneh method (Fig. 7), there is no fundamental reason of why each row of observations should predict the maturity variable since this is explained by the correlation of the process variables in time, and not from an individual time sample in the batch (I guess this is why the PLS model gives multiple solutions for a certain predicted time).

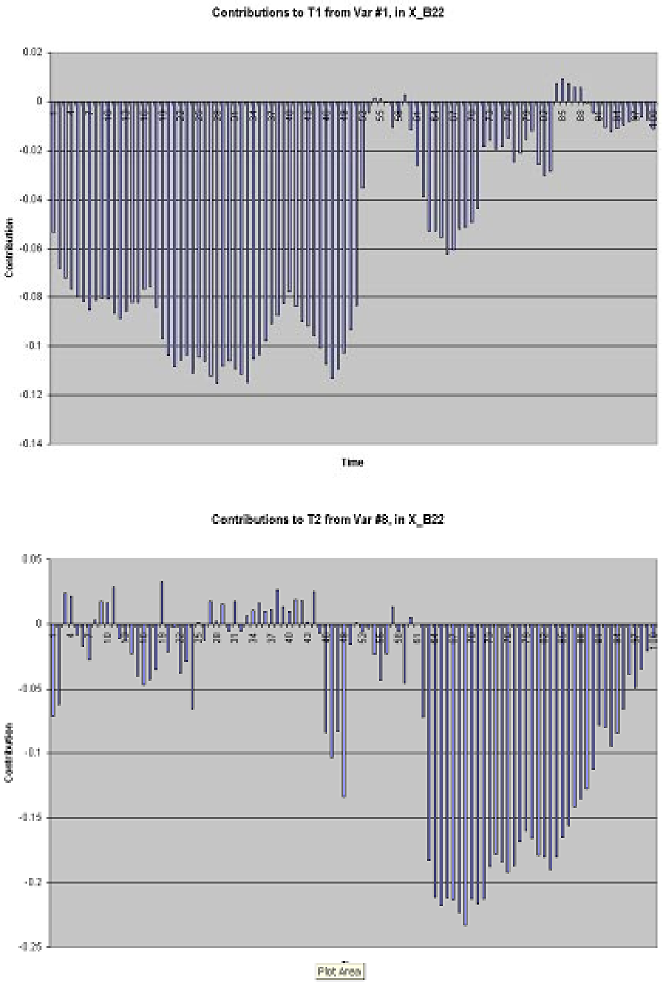


Fig. 21. Contributions to the scores for two different variables in batch 22

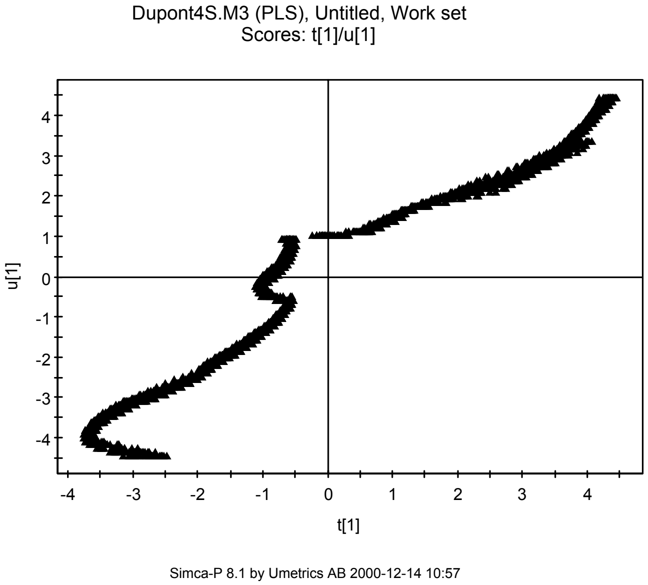


Fig. 22 t1u1 plot for the PLS model done in SIMCA-P

### *B. Handling missing data when monitoring new batches.*

Mentioned earlier, there are three different approaches to solve the problem of missing data when monitoring a new batch (Fig. 5ii). One of the approaches suggested by Nomikos and MacGregor include the use of the ability that PCA has to handle missing data.

This approach is the most realistic way of handling missing data because of the following: the ability that PCA has to handle missing data does not fill the blanks with any value, but it considers that value as non-existent; therefore, the score value is not biased with any projection about the future, like filling with zeros or with current deviation or with estimated values from an ARIMA model. This is fundamental, treat missing as missing, if we want to predict the future observations of a batch, we will require a mechanistic model in order to achieve enough robustness and reliability.

A mechanistic model may take much time to build and, from a business perspective, can be very expensive. So, a rigorous solution to the problem is presented, according to the suggestions of Nomikos and MacGregor.

The rigorous solution to this problem is to build as many models as time samples in the batch, this solution was not considered before because of the computational effort that it could mean in a real time application. With the actual computing power available, this not a problem any more.

The model "n" will include time samples from 1 to "n" for all variables, so, when a new batch comes, one different model will be used each time a new sample comes, in this scheme there is absolutely no assumptions about the future values of the variables, because they are not considered at all in the model.

The nylon data was used, and 100 models were built with the known 36 good batches, the rest of the batches were monitored using the 100 models in order to calculate the Hotelling T2 and the DModX for each time sample.

In general, we could appreciate that the alarms given when the missing data ability from PCA was used, were very similar to the alarms given by the rigorous solution. These last ones kept the alarm state for more time (Fig 23-26).

There is concern about the robustness of missing data ability of PCA monitoring scheme in the beginning of the batch, this particular study showed that it is robust enough. Building as many models as time samples does take a big effort and time, but for some cases it is worth the job. An online application with the rigorous solution is feasible thanks to the features and speed of some solutions available in the market, and of course, a hybrid solution always could be implemented.

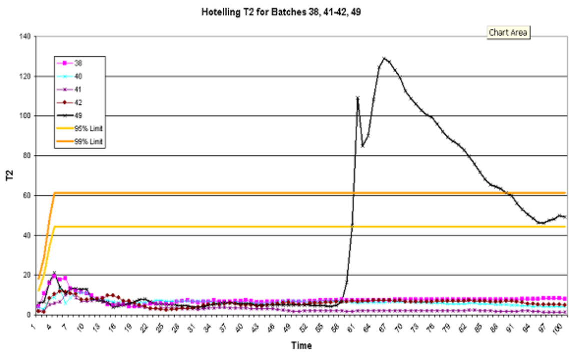


Fig. 23. Rigorous monitoring shows alarm for batch 49

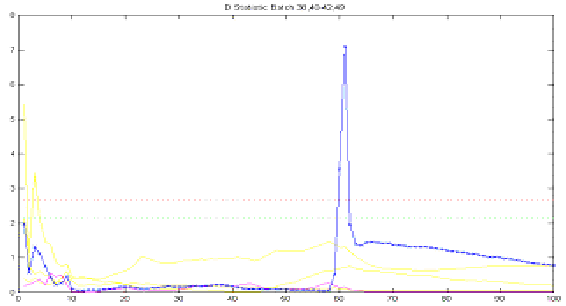


Fig. 24 Monitoring using missing data ability from PCA shows alarm for batch 49.

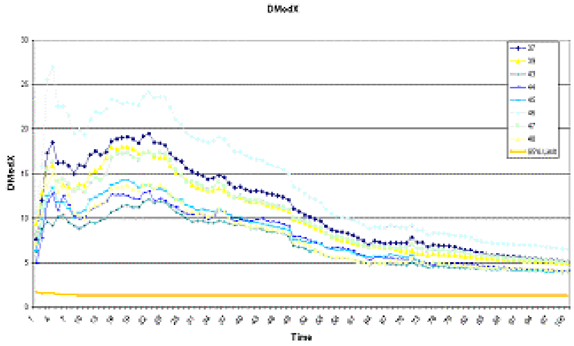


Fig.25 Rigorous monitoring shows faulty batches from the beginning

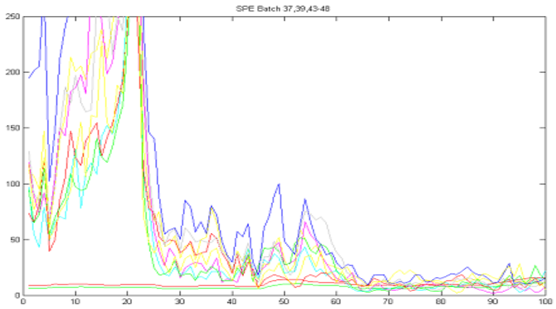


Fig. 26 Monitoring using missing data ability from PCA shows alarms for batches from the beginning

*C. Data from a PVC process.*

This data set consists of 37 batches each one has a different duration, and 6 variables are sampled. This is a suspension polymerization process. The variables are:

1. Temperature of cooling water at jacket outlet
2. Temperature of cooling water at jacket inlet
3. Cooling Water Flow rate thru the jacket
4. Temperature inside the reactor
5. Pressure inside the reactor
6. Amperes used by the agitator system

The first task is to align the batches to avoid the different length problem. The aligning was done using the percent completed of total time; the intermediate values were interpolated since each variable has a different sampling time.

The first model, built with the raw data shows some clustering in the t1t2 space of BatchSPC (Fig. 27), while in the control charts for SIMCA-P there is no imminent clustering (Fig 28), from this point, SIMCA-P is no longer used for this particular study because it will not provide the necessary resolution to analyze the data set.

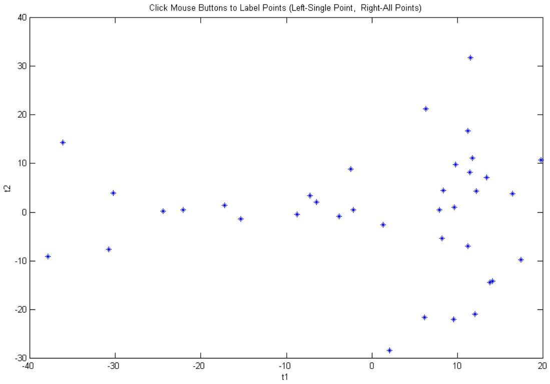


Fig. 27. Clustering in BatchSPC

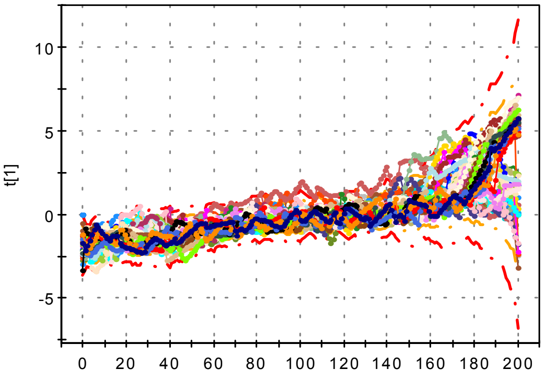


Fig 28. SIMCA-P Control Chart for the PVC raw data

In this work, we have posed that a principal component is a reflection of the fundamental phenomena occurring in the batch, hence, to reinforce the data, one extra variable was included to account the total energy removed by the cooling system from the reactor, this will be variable #7.

After including this new variable, the clustering in the data is even stronger (Fig. 29). According with the data owner, batches 1 to 9 are from the cold season of the year; the rest is from the warm season of the year.

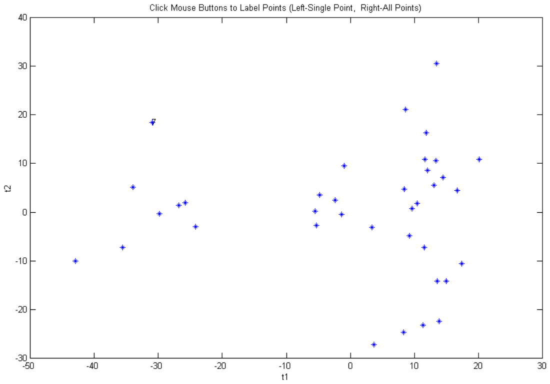


Fig.29 Clustering when energy balance is included

The following to do is to look at the differences between the clusters in the contributions (Fig. 30) we can see that the difference between batches from cold season, and batches from warm season is the flow thru the jacket, and the total energy removed from the reactor, this is precisely what was expected from the model.

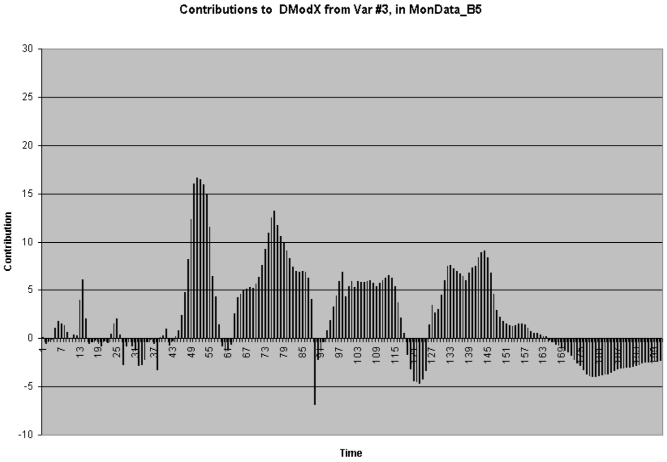


Fig. 30 Heavy contribution from the water flow thru the jacket

Once we know that the difference between clusters is something by nature in the system, we build a model for each set of batches, one PCA model for batches 1 to 9 and another one for batches 10 to 37 (Fig 31,32).

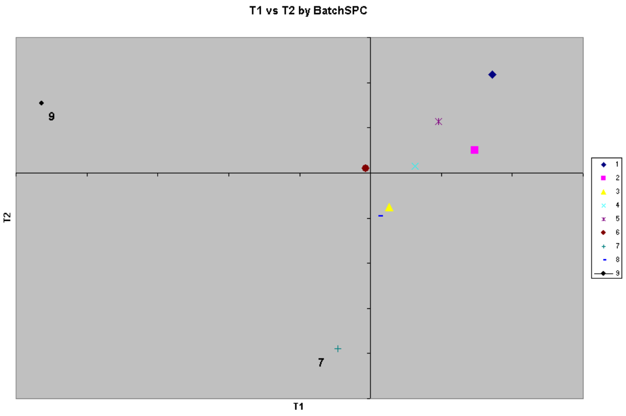


Fig. 31 PCA for cold season batches has two outliers, batch 7 and 9

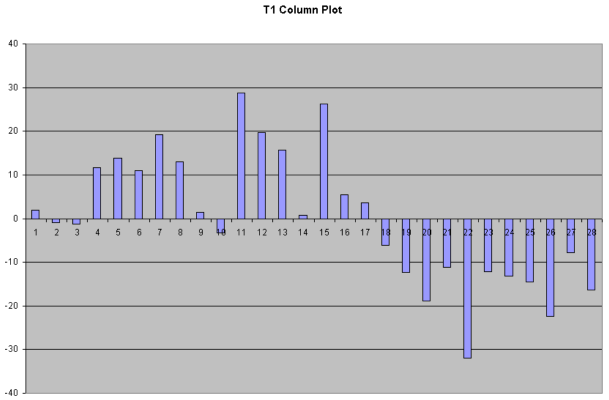


Fig. 32 T1 for warm season batches has three batches strongly different

For each set of batches, the contributions helped to find out the specific variables driving the batch away from the expected.

For batches 7 and 9, the control system had upsets in the water flow, for batches 13, 20 and 31 control upsets and different temperatures in the cooling water made them different from the rest.

## V. PLS USED TO INFERR BATCH PROPERTIES (SOFT SENSING)

PLS can be used for inferential purposes, most polymer processes have to wait some time while the lab is giving the quality analysis of the product and, in some cases, the samples are not even taken batch to batch. In other batch processes, the products from several reactors go to a blending operation or into the same separation process (in order to recover solvents or non reacted raw material) so it is important to determine if the product from a certain batch is within the quality specs, to avoid it the spoilage of other products, and big economic impact.

PLS gives the tool for building inferential models, and therefore we could predict in real-time the properties of certain batch, and decide if it is good product or not. It is important to say that a quality property cannot always be predicted with a mechanistic model.

In this case we calculated the total reaction time for the 37 batches from the PVC data set (recalling that all the batches had different time lengths). These times were used as a Y variable and a PLS model was built in order to predict the total batch time.

Fig. 33 shows the results given by the model, the predictions are actually very good, and the worst prediction only differs in 3% from the observed total batch time. Using the raw data and the raw data with the extra energy variable improves the prediction in 4.98%, this is not significant because it means that the worst case improves only 36 seconds closer to the observed value (the difference between observed and predicted goes from 12.049 to 11.447), Fig 34 shows t1 vs. u1 for this model and a clear linear relationship can be observed.

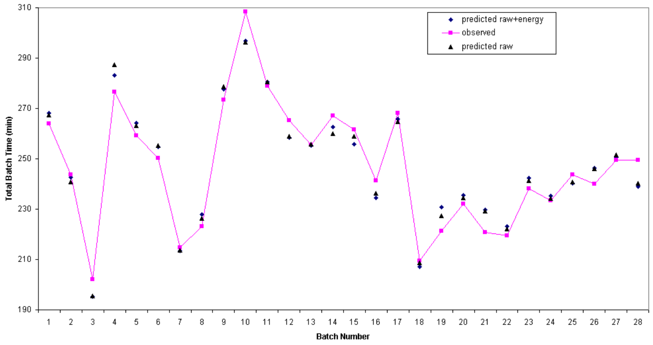


Fig. 33 Predicted vs. observed total batch times.

Analyzing the loadings for this PLS model we see that the heavy variables are the pressure inside the reactor and the temperature of the cooling water into the system, these two variables account for the biggest variation of the data in Y, once again, statistic and deterministic approaches find a common space to be applied when interpreting the loadings of the model, the thermodynamic in the reactor are strongly correlated with the total reaction time, and this is why the PLS model is so good.

Another study was done to see if the PLS prediction was maintained during the batch. Fig 35 shows the difference between the observed total batch time and the prediction during the batch (the missing data was handled as missing data in the PLS model) we can clearly see that the prediction is very bad with partial batch data, and since there is no fundamental reason for a partial batch to be correlated with the total batch duration, then we again obtain what we expected.

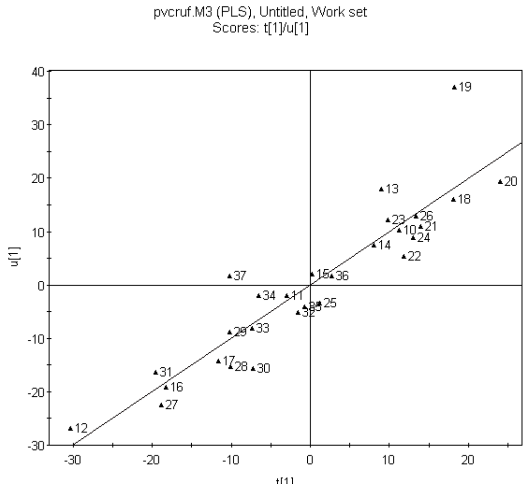


Fig. 34. t1 vs. u1 for the PLS model predicting total batch time.

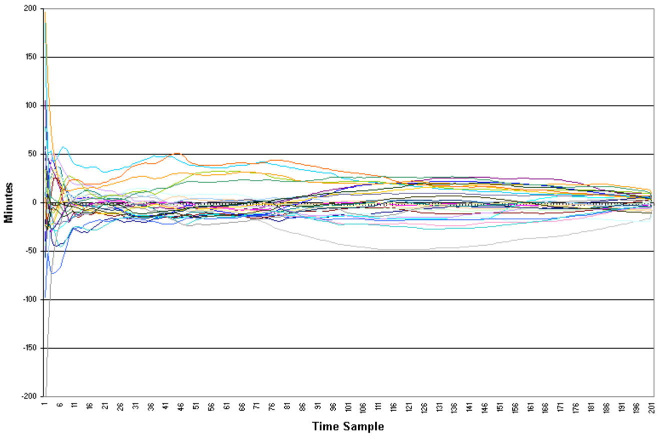


Fig. 35. Difference between observed and predicted along the batch duration

## VI. BUILDING A BUSSINESS CASE

Any industrial analysis should have a business case attached, or an economic impact study. PCA and PLS gives us a powerful tool to analyze the opportunity operational cost of a certain process.

When we build a PCA/PLS model around certain process data, the principal components will define the principal sources of variation, studying the contributions to the scores and depending on the nature of the process we can certainly calculate the economic impact of the process when it is located in certain regions of the score scatter plot.

This is particularly useful when we deal with big processes where almost all the operations are correlated ones with others. PCA can be a powerful tool to find sources of economical variation (because of quality, because of operational costs, raw material quality, seasonal impacts, etc).

For example, in the PVC data, the t1 t2 scatter plot shows seasonal clusters (Fig. 29) the contributions to the scores show that the main difference between clusters is the water flow thru the jacket.

From here we can calculate the differences between clusters. The cold season batches use 3 $m^3$ less of water per batch, from this point we could easily build a business case or an economic analysis.

## VII. CONCLUSIONS AND FUTURE WORK

From this work we have learned that time wise is the most reasonable way of unfolding the 3D data matrix from industrial batch data. For monitoring new batches, using the ability that PCA has to handle missing data is the most realistic approach to solve this problem; implementing a rigorous solution, building as many models as time samples, gives the same alarms for batch upsets.

A PLS model has to be supported and interpreted by fundamental laws and phenomena occurring in the batch. A well supported PLS model can be successfully used to predict batch/product properties.

Last, the best model is not the one that captures 100% of the variance of a certain data set; the best model is the one that captures the relevant and fundamental variance of a certain data set.

Future work in this area includes the implementation of a real time batch-aligning algorithm like DTW and the development of an adaptive scheme for batch MPCA/MPLS.

## ACKNOWLEDGMENTS

Special thanks to Dr. John F. MacGregor for his patience and guidance, Dr. Dora Kourti for the BatchSPC application and the nylon batch data, Dr. Svante Wold for his support using SIMCA-P, and Mr. Lino Campos for the PVC data used in this work.

## REFERENCES

1. J.V. Kresta, J.F. MacGregor, T.E. Marlin, ‘Multivariate statistical monitoring of process operating performance’, *Can.J.Chem.Eng*. **69** (1991) 35-47

2. S.Wold, P. Geladi, K. Esbensen, J. Öhman, ‘Multi-way principal components and PLS-analysis’, *J. Chemometrics*. **1** (1987) 47-56

3. P. Nomikos and J.F. MacGregor, ‘Monitoring of batch processes using multiway principal component analysis’, *AIChE J.* **40**, (1994), 1361-1375

4. P. Nomikos and J.F. MacGregor, ‘Multi-way partial least squares in monitoring batch processes’, *Chemometrics Intell. Lab. Syst.* **30**, (1995), 97-108

5. P. Nomikos and J.F. MacGregor, ‘Multivariate SPC charts for monitoring batch processes’, *Technometrics*, **37**, (1995b), 41-59

6. S. Wold, N. Kettaneh, H. Friden and A. Holmberg, ‘Modelling and diagnosis of batch processes and analogous kinetic experiments’, *Chemometrics Intell. Lab. Syst.*, **44**, (1998) 331-340

7. A. Kassidas, J.F. MacGregor and P.A. Taylor, ‘Synchronization of batch trajectories using dynamic time warping’, *AIChE J*. **44**. (1998) 864-875.