2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

Application of multivariate batch data analysis for troubleshooting of end-point product quality

Zdravko Stefanov and Leo Chiang, *The Dow Chemical Company*

Abstract—**A product is produced in a production train that consists of two batch reactors in series. The quality specifications require that a particular critical impurity (CI) in the product should be below 0.4%. The plant was experiencing series of poor quality batches with much larger critical impurity amounts and that was a cause of significant trouble.**

Multivariate batch data analysis discovered that possible mixing issues in the second batch reactor (R2) were the cause of the problem. Therefore the recipe was changed in a way that improved the mixing.

Before the proposed solution was implemented, 18 out of specification batches were produced, out of total of 20 batches which is 90% out of specification. After the solution was implemented, only 3 out of specification batches were produced, out of total of 98 batches, which is 3% out of specification.

The results achieved based on this work allowed the plant to produce the product within specifications and maintain shipping schedule.

INTRODUCTION

THE product that is discussed in this paper is produced in two batch reactors in series. The first reactor performs a preparation step where most of the reagents are loaded and the first reaction step occurs. The second reactor produces the final product. The reactors are multistep, i.e., there are multiple sequential steps that are followed for each batch in each reactor. The step sequences for both reactors are given in TABLE I.

The end-point batch quality depends on the level of a critical impurity (CI) which is a side reaction product. The quality specifications do not allow this impurity to exceed 0.4% in the final product. At the time multivariate data analysis support was required the plant had made multiple batches with impurity levels much higher than the specification. To troubleshoot the plant, three categories of data analysis were performed.

Manuscript received September 22, 2010. Zdravko Stefanov is with the Dow Chemical Company, Analytical Technology Center, Freeport, TX 77541 USA (phone: 979-238-5357; fax: 979-238-0336; e-mail: zstefanov@dow.com).
Leo Chiang is with the Dow Chemical Company, Analytical Technology Center, Freeport, TX 77541 USA (e-mail: hchiang@dow.com).

The first analysis is of Data Set A that includes analytical measurements of different properties of the final product and some batch initial conditions. It does not include process data from the time trajectories of the batches. The second analysis is performed on two sets of process data that do include the time trajectories of the batches (Data Sets B and C). The third data analysis is of Data Set D, a very small set that includes only one process variable plus the %CI for batches made in a subsequent campaign and it is used for final validation of the previous findings.

The main point of the paper is to show the applicability and usefulness of multivariate batch data analysis for troubleshooting of batch processes. Therefore the mathematical background of the multivariate data analysis is not included in this paper due to limited space.

TABLE I
REACTOR SEQUENCES

Step number	Step description
Reactor 1 (R1)	
3	Load reactant 0
4	Load reactant 3
5	Load reactant 2
6	Load reactant 1
7	Heating up
8	Reacting
9	Cooling down
10	Waiting for R2
11	Prepare transfer line
12	Transfer to R2
Reactor 2 (R2)	
33	Receiving material from R1
34	Load reactant 6
35	Load recycle material
36	Load reactant 7
37	Load reactant 5
38	Reacting
39	Adjust property 1 and wait for next unit operation
40	Transfer to next unit operation

ANALYSIS OF ANALYTICAL AND BATCH INITIAL CONDITIONS DATA (DATA SET A)

The data set includes the amounts of reactants 1, 2, 3, 5, 6 and 7; five analytical measurements of side reaction products and the maximum temperature during step 38 in R2 (the X variables). The Y variable is the concentration of the critical impurity (%CI). The set contains 25 batches. The variables in this set were proposed by the plant subject matter experts as the data were already available and any found correlations could provide insight that might help to solve the problem, while more complete plant data set that includes the time trajectories of the process variables could be collected. This data set contains values measured once per batch and batch conditions. Therefore it was not necessary to perform batch unfolding.

The data set was analyzed using Partial Least Squares (PLS) [1]. Only three variables from the data set were included in the model – the amount of reactant 1 and two of the analytical measurements of side reaction products. The model parameters are given in TABLE II.

TABLE II
DATA SET A PLS MODEL PARAMETERS

Number of variables	Number of principal components	R^2X	R^2Y	Q^2Y
3	1	0.605	0.665	0.583

The model explains about 67% of the variation in the CI concentration. The cross-validation R^2 (Q^2Y) is close to the R^2. The model fit is given in Fig. 1. It is observed that the trend in the CI concentration variation is followed well, even if after batch 13 in the data set the predictions are poorer compared to the first 13 batches.

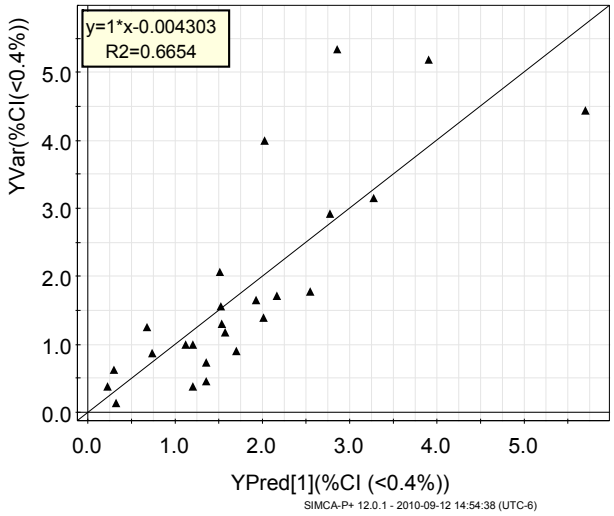

Fig. 1. Observed versus predicted %CI content, data set A PLS model

The analysis shows that the amount of reactant 1 has negative correlation to the %CI and the two analytical measurements of side reaction products have positive correlation to the %CI. The correlation between the amount of reactant 1 and the %CI is weak, and if the amount of reactant 1 is removed, the model R^2 drops only to 0.62, therefore it can be concluded that only the two analytical measurements of side reaction products are of importance. The positive correlation with one of the side product's concentration could be explained by the chemistry of CI formation. High residual of the side reaction products could mean that the active complex formation with reactant 7 was somehow hindered, and therefore the chemicals could not react fully to the final product. This will leave more precursor to react to CI. The side products' concentrations are correlated. Because there is only one major source of variation, the model contains one principal component.

Regarding the decrease in model quality for the batches after batch 13, it was decided to split the data in two parts, one from batches 1 to 13 and another batches 14 to 25. The predictions are much better when the data are separated, see Fig. 2.

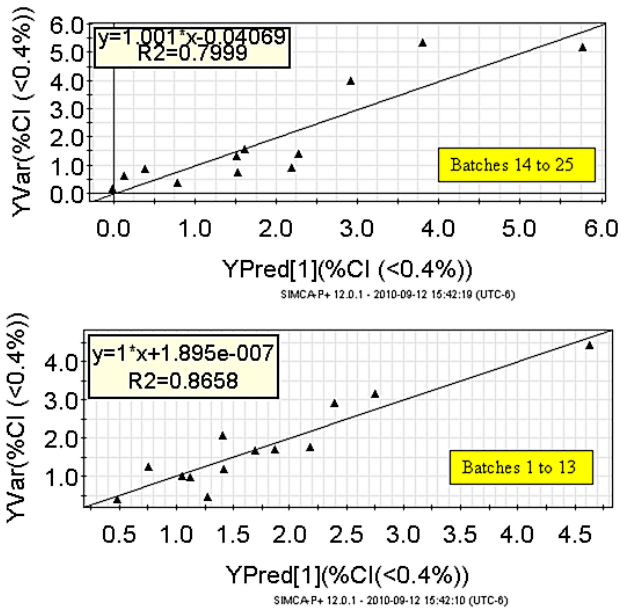

Fig. 2. Observed versus predicted %CI for two subsets of batches, data set A

This improvement in prediction was assumed to be due to another process factor not included in the dataset. After the plant subject matter experts were consulted about what possibly changed at batch 14, it became clear that a new reactant 7 batch was used for the batches from 14 onward. The change in reactant 7 seems to be the contributor to the change in the behavior of the plant.

PROCESS DATA ANALYSIS

The results presented in the previous section make sense from process knowledge; however, there is no particular handle (i.e., input variable) that could be used to drive the %CI back into specification. All variables identified as correlated to the %CI in Section II are analytical results after the product is made. Therefore process data were collected to perform a full batch data analysis, meaning that the entire batch trajectories of the process variables during each batch were used as inputs (X). The batch data were unfolded from 3 dimensions (batch, time and process variables) to two dimensions using the scheme given in Fig. 3.

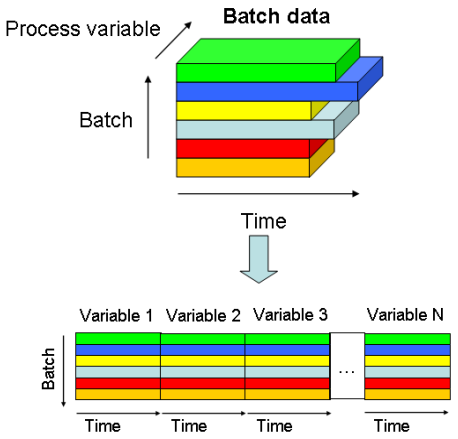

Fig. 3. Batch data unfolding scheme

The batch unfolding scheme in Fig. 3 is otherwise known as Nomikos and MacGregor [2] unfolding and is very effective in detecting batch-to-batch differences, which is the case here.

A. Analysis of 35 batches made in March 2008 (data set B)

This data set includes a total of 29 process variables, the %CI quality variable and includes all steps in R1 and R2. The objective here is to build a PLS model to predict the %CI. If a good model is built, then the contributing variables can be investigated and a possible handle to drive the %CI back into specification can be found. In this case, indeed, a good PLS model was built. The model predicts very well the batches included in the data set, see Fig. 4. The model parameters are given in TABLE III.

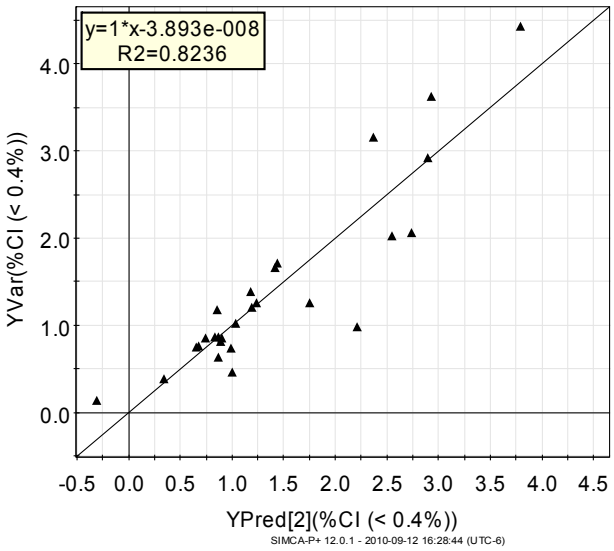

Fig. 4. Observed versus predicted %CI, data set B PLS model

TABLE III
DATA SET A PLS MODEL PARAMETERS

Number of unfolded variables	Number of principal components	R^2X	R^2Y	Q^2Y
77	2	0.733	0.824	0.746

The actual process variables used in the model are:

- R2 agitator amps during step 34
- R2 weight during step 34
- Duration of step 6 in R1
- Duration of step 36 in R2

The first two variables reflect the amount of work put by the agitator in R2, respectively the mixing quality. More mixing work delivered (amps of the agitator higher) corresponds to lower %CI. This observation is also confirmed by past process experience. The plant subject matter experts shared that the mass transfer during the reaction is running on 'the ragged edge' and if the process tips over the edge, dire CI results can occur. The correlation of the %CI to the R2 agitator amps during step 34 is shown in Fig. 5. (Note that in the plot, the variables are scaled to the range of 0-1.) It is clear that there is good inverse correlation between these variables.

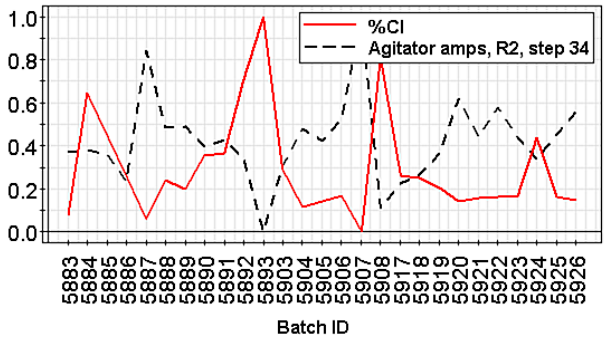

Fig. 5. Relation between R2 agitator amps in step 34 and %CI

Using the findings from the modeling of this data set and after a discussion with the plant subject matter experts, it was decided to take action and change the future batches in two ways.

– First, the batches were made "skinnier", i.e. the reactant 1 amount was reduced by 10% compared to the original recipe.
– Second, the agitator in R2 was repaired after some mechanical issues were discovered.

After these changes were implemented, the plant resumed operation and all consecutive batches of the product were produced in specification.

Although the implementation of these measures led to successfully finishing the campaign, more analysis was performed on the data from the entire campaign. The main reason for this was that the "skinny" batches led to decrease in throughput and the plant slowly increased the recipe to normal reactant 1 charge. However the %CI did not increase. Therefore it was of great interest to investigate this and check if there are some additional factors that might have influenced the CI formation. This analysis is covered in the next section.

B. Analysis of 88 batches made from March and April 2008 (data set C)

1) Analysis of the large variation in the %CI

The process variables in this data set are the same as these in data set B. Again a good PLS model was built. The model predicts very well the batches included in the data set, see Fig. 6. The PLS model parameters are given in TABLE IV. The time series plot in Fig. 7 shows that the model captures the big changes in the %CI. A very indicative observation is batch 5999, where a single bad batch is predicted well.

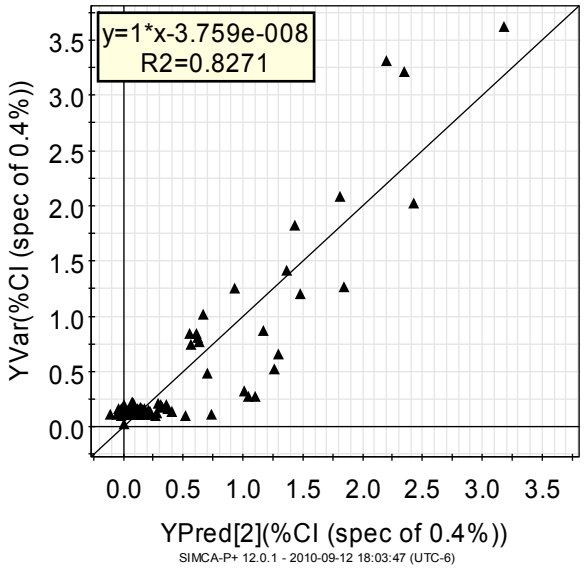

Fig. 6. Observed versus predicted %CI content, data set C PLS model

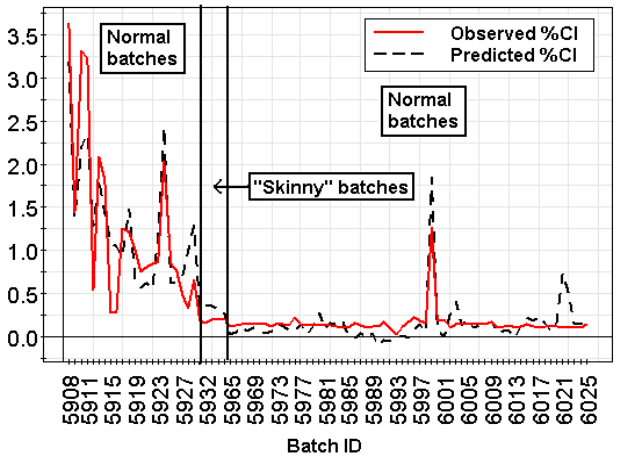

Fig. 7. Observed versus predicted %CI content, data set C PLS model, time series

TABLE IV
DATA SET C PLS MODEL PARAMETERS

Number of unfolded variables	Number of principal components	R^2X	R^2Y	Q^2Y
602	2	0.601	0.827	0.675

The main question to answer here is why the difference before and after the "skinny" batches. The contributions to the predictions reveal that there are two main contributing variables:

– R2 agitator amps during step 34
– R2 parameter 1 during step 40

The importance of the R2 agitator amps in step 34 is not a huge surprise, since it was discovered to be important in the previous analysis. If this variable is plotted together with the %CI (Fig. 8, note that all variables are scaled to the range of 0-1), it is clear that the amps were even higher after the "skinny" batches. It is also observed that the agitator amps clearly indicated the high %CI in batch 5999.

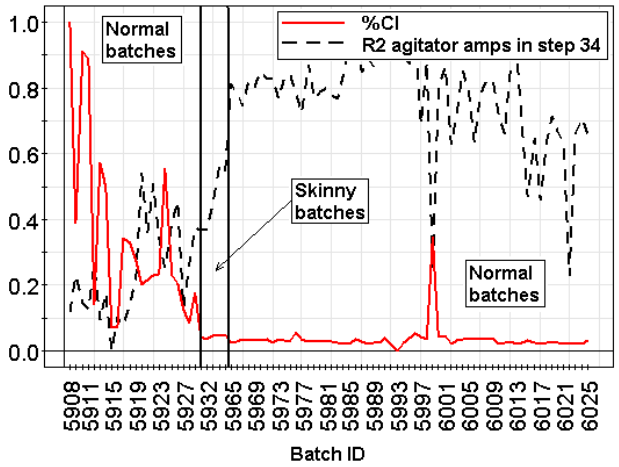

Fig. 8. %CI versus R2 agitator amps in step 34, data set C

The other variable that experiences significant change after the "skinny" batches is the R2 parameter 1 in step 40. If this variable is plotted together with the %CI (Fig. 9, note that the variables are scaled to the range of 0-1), it is clear that the parameter 1 dropped quite significantly after the "skinny" batches. It shows that the parameter 1 variable can be another important indicator for the CI formation.

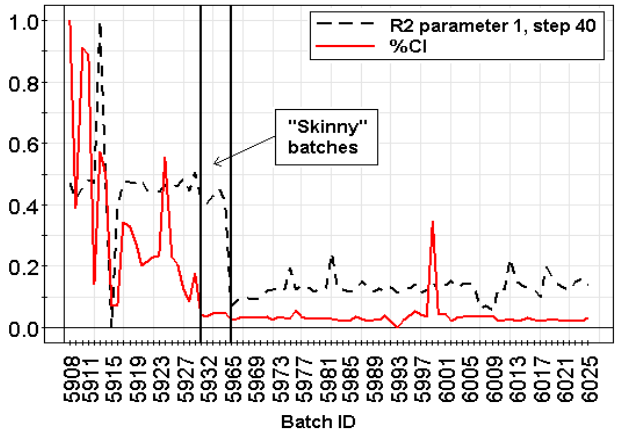

Fig. 9. %CI versus R2 parameter 1 in step 40, data set C

2) Analysis of the small variation in the %CI

Many batches in this campaign were produced well within specification. It was interesting to see if for these batches the process experience from past campaigns still holds. For that purpose only these batches were modeled and the important variables investigated.

The first model built using this data predicts the %CI well, except for two batches, 5976 and 5993. The score plot does not indicate anything unusual for these batches. Also, the distance to the model plane (SPE) of these batches is also within the critical limits. These batches are therefore unexplained outliers and the error could be due to analytical issues or some other factors not reflected in this data set. Therefore these two batches were removed. Consequently, more outlying batches were found, that were also removed, and all removed batches represent 14% of the original number of observations. This means that 86% of the original data was retained to build the final model. The final model predicts the %CI very well, see Fig. 10. The model parameters are given in TABLE V.

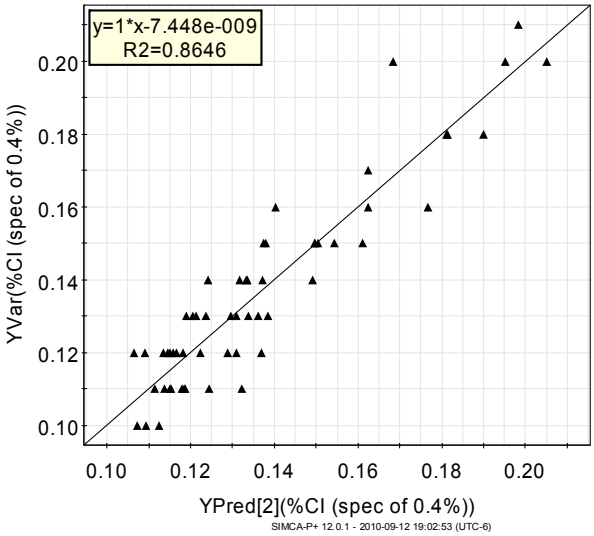

Fig. 10. Observed versus predicted %CI, low %CI batches, outliers removed

TABLE V
DATA SET D PLS MODEL PARAMETERS, LOW %CI BATCHES

Number of unfolded variables	Number of principal components	R^2X	R^2Y	Q^2Y
10	2	0.930	0.865	0.850

The two process variables used to build the model are the duration of step 8 in R1 and the R2 parameter 1 in step 38, minutes 2 to 10 in the step. The parameter1 in step 38 alone explains 70% of the variation in the %CI, see Fig. **11**. This observation was confirmed by the plant subject matter experts and it was pointed out that this behavior is expected when the %CI is within its normal range.

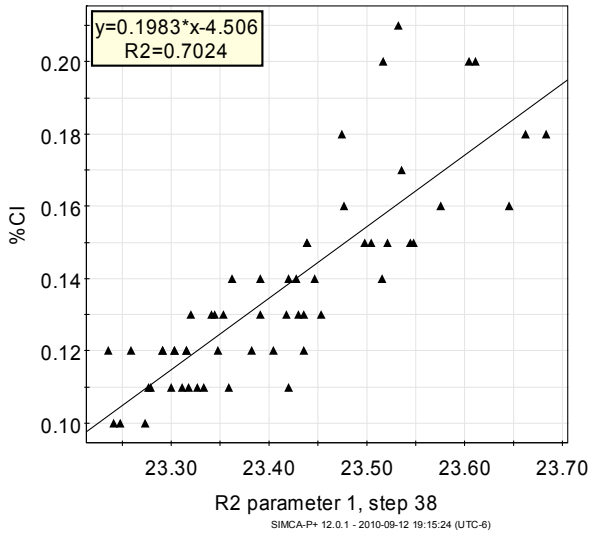

Fig. 11. %CI versus R2 parameter 1 in step 38, low %CI batches

C. Analysis of 23 batches made in July 2008 (data set D)

For this campaign, no multivariate data analysis was performed to model the %CI. Instead, the R2 agitator amps were monitored to detect possible problems. The campaign started relatively well and six batches of good product were produced with %CI in specification. Then the CI concentration increased and after a series of poorly performing batches, the plant applied the "skinny" recipe that was found to work well in the spring campaign. That action seemed to fix the problem and the rest of the batches were produced within specification. With respect to the correlation between the %CI and the agitator amps in R2 in step 34, it was still present, as it is clear from Fig. 12.

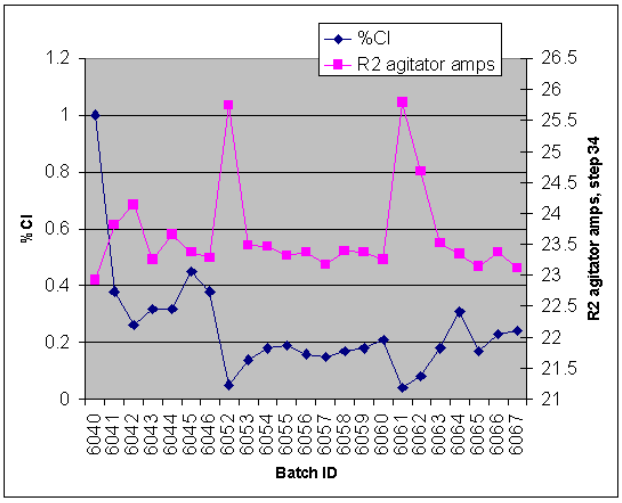

Fig. 12. %CI versus R2 agitator amps in step 34, July 2008 campaign

Conclusions

In summary, the analysis of the March and July campaigns revealed that the concentration of the CI in the product depends greatly on the mixing efficiency in R2. Both observations are confirmed by past plant experience, however now there is a very good indicator for possible troubles with the CI. That is the R2 agitator amps in step 34. This variable consistently correlates with the CI concentration in the product. Another result of this investigation is the discovery of the "skinny" recipe, which works fine with respect to the quality of the product.

The agitator amps in step 34 are an indicator, not a handle that can be manipulated to drive the %CI back in specification. The negative side of the "skinny" recipe solution to the CI problem is obviously the decrease in the throughput, which is a very undesirable side effect. Therefore more fundamental research is required to discover a better handle to control the CI impurity. Nevertheless, a handle is better than none and the multivariate batch data analysis was the technology that made the critical difference and helped to discover the handle of the "skinny" recipes. This is a unique situation in batch processing, where run-to-run control of the same product is achieved by using of different recipes. Usually different recipes are used for making of different products. In this case different recipes produce different levels of the CI. If an unmeasured disturbance affects the process and the batches start drifting towards higher %CI, the recipe is changed to a "skinnier" one until the %CI is back within specifications.

References

[1] H. Martens and T. Naes, "Multivariate calibration, New York: John Wiley and Sons, 1991.

[2] P. Nomikos and J.F. MacGregor, "Multi-way partial least squares in monitoring batch processes," *Chemometrics and Intelligent Laboratory Systems*, vol. 30, pp. 97–108, 1995.