MULTIVARIATE ANALYSIS FOR QUALITY IMPROVEMENT OF AN INDUSTRIAL FERMENTATION PROCESS

Leo Chiang[*a], Arthur Kordon[a], Lawrence Chew[b], Duncan Coffey[a], Robert Waldron[a], Torben Bruck[b], Keith Haney[b], Annika Jenings[b], and Hank Talbot[b]

[a] *The Dow Chemical Company, Corporate R&D, 2301 Brazosport Blvd., Freeport, TX 77541*
[b] *The Dow Chemical Company, Biotechnology, 5501 Oberlin Dr., San Diego, CA 92121*

Abstract: The results of a successful implementation of multivariate analysis on multiple fermentations are shown in the paper. To minimize batch-to-batch variability of an industrial fermentation process, multi-way partial least squares (PLS) was used. Eighteen baseline batches from different fermentors were analyzed. After applying diagnostics tools and contribution charts, three batches were clearly identified to be abnormal, and even among the remaining batches, inconsistencies were found. The root causes of the variability were determined by a combination of variable importance in the projection plot and fermentation knowledge. A new fermentation procedure was applied and the quality improvement was demonstrated on 20 new batches. These batches were more consistent as evidenced by the improvement in the model fit ($R^2X = 0.829$ for the new batches versus $R^2X = 0.681$ for the baseline batches) and in the percent of out-of-specification (3.3% for the new batches versus 20.4% for the baseline batches). The on-line multi-way PLS model was shown to detect bad batches promptly and to determine abnormal variables accurately. The success of this implementation demonstrates the value of applying multivariate analysis to large-scale industrial batch bioprocesses.

Keywords: Multivariable systems, fault detection, batch control, fermentation processes, and biotechnology

1. INTRODUCTION

To achieve consistent product quality from a batch process, minimizing batch-to-batch variability is important. Multivariate statistical techniques such as principal component analysis (PCA) or partial least squares (PLS) are useful for quality improvement (Chiang *et al*., 2001; Beebe *et al*., 1998; Zhang and Lennox, 2004). In off-line applications, these techniques can identify and pinpoint the root causes of batch-to-batch variability. In on-line applications, PCA/PLS models are used to monitor batch conditions. The objective is to identify and correct abnormal conditions early enough to avoid out-of-specification product.

In this paper multi-way PLS analysis was used for quality improvement of an industrial fermentation process at the San Diego biotech facility of The Dow Chemical Company. This application illustrates the use of multivariate analysis for solving typical industrial problems where root causes are unknown. This paper also addresses practical issues of industrial data analysis and outlines the steps needed to minimize variability in the final product quality.

2. METHODS

2.1 Multi-way PLS

Applying multi-way analysis for batch data is a two-step process. The first step is to unfold the three-way batch data into two-way data (see Figure 1) and the second step is to apply regular PLS/PCA analysis.

In this mode of unfolding, each column contains a particular variable over all time periods for all batch runs and each row contains all variables at a particular time for a particular batch. After unfolding, regular PLS is applied with the X block (independent variables) as the unfolded matrix and the Y block (dependent variable) as the maturity variable (Wold *et al*., 1998). A monotonically increasing variable that is related to percent completion of a batch is used as maturity variable. It is preferable to use a quality variable. However, for the case when the uncertainty in the quality variable is too high, elapsed time can be used as maturity variable instead.

This way of unfolding is attractive for on-line implementations because the score calculation and maturity prediction can be computed directly based

[*] To whom all correspondence should be addressed. E-mail: hchiang@dow.com

on on-line measurements. Contribution charts can be easily computed and interpreted. Similar to regular PLS, the T^2 and Q statistics are used to monitor the process, while the predicted Y variable gives information about the maturity of a batch. Because of these advantages, this mode of unfolding was used.

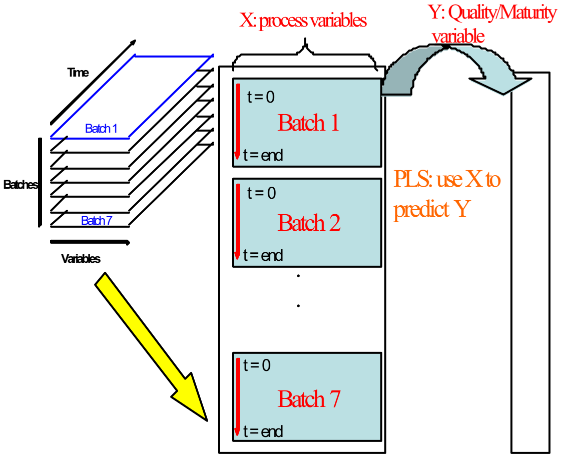

Fig. 1. Illustration of multi-way PLS for batch data.

2.2 Fermentation process

The multivariable analysis was performed on experimental batches from the Dow Chemical San Diego facility. The scope of the investigation included six 20 L fermentors. The key objective was to eliminate problematic fermentation issues that prevent 100% fermentation success rate. The results from this project will be leveraged to other large-scale bioreactors.

Fermentation consists of two phases: growth and production. To enhance the sensitivity on analyzing batch data, separate PLS models have been applied for each phase using the SIMCA-P software (Umetrics, 2003). In SIMCA-P, non-linear iterative partial least squares (NIPALS) is used in PLS computation and leave 1/7 out cross validation is used to determine the optimal number of PLS components. In the growth phase, X-block contains on-line process variables of interest and Y-block contains the optical density (OD), which measures the dynamics and the maturity of cell growth. In the production phase, X-block contains the same variables as in the growth phase, but Y-block contains relative activity/protein yield (referred to as activity in this paper). Quality decisions are made based on OD and activity. Therefore, it is more meaningful to use these two variables, instead of elapsed time, in the Y block.

The purposes of applying PLS for analyzing the batch data are:

- Identify batch-to-batch consistency
- Identify key variables that correlate with the OD and activity
- Understand the root causes of bad batches

Eighteen batches were collected during a month period to identify the baseline performance of the process. Process data were measured on-line every minute, while quality variables (OD, activity, etc.) were measured on an hourly basis. Because OD and activity were monotonically increasing, all missing values between sample points for the Y-block were linearly interpolated using SIMCA-P.

3. RESULTS AND DISCUSSION

3.1 Consistency for the baseline batches

Figure 2 plots the predicted against the observed OD for the 18 baseline batches in the growth phase. The model fit ($R^2X = 0.816$ and $R^2Y = 0.980$ for a PLS model with 6 components) is reasonably good, indicating that some process variables correlate well with the OD. This is in spite of the fact that OD and elapsed time for each batch vary at the end of the growth phase.

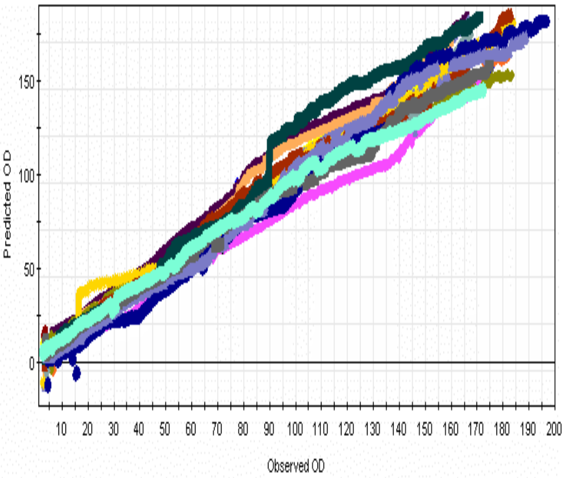

Fig. 2. Predicted OD versus observed OD for the baseline batches in the growth phase.

From a control standpoint, it is important to determine the key variables that are related to the OD. To accomplish this, the variable importance in the projection (VIP) plot is used (Eriksson *et al.*, 2001). The VIP plot, computed based on the PLS weights and the variability explained, ranks the process variables in terms of their relative contribution in predicting OD.

As shown in Figure 3, the top variables in the VIP plot have higher correlation with the OD. On the average, variables with magnitude greater than one in the plot are more relevant in predicting OD. Poor control in these variables will definitely result in a large inconsistency in the OD. At the same time, variables with low correlation to OD must not be ignored. If these variables are related to the key process parameters (e.g., through control loop), poor control of these variables will affect the key variables. This will have a cascade effect on the consistency of the OD. The VIP plot prioritizes the effort to determine the root causes of batch-to-batch inconsistency.

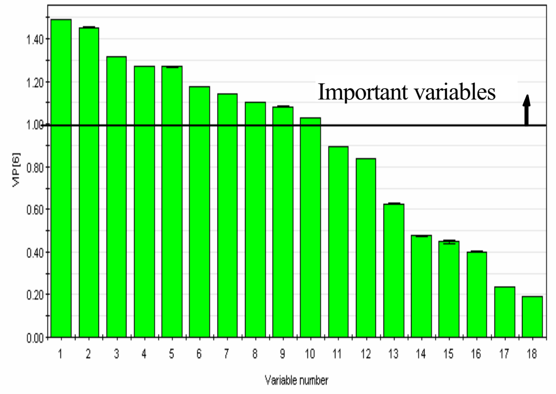

Fig. 3. The VIP plot for the baseline batches in the growth phase.

Figure 4 plots the predicted against the observed activity for the production phase. Comparison between production and growth phases shows that the model fit is worse in production phase ($R^2X = 0.642$ and $R^2Y = 0.848$ for a PLS model with 6 components). This indicates that correlation between the process variables and activity is weak for some batches and the correlation is not consistent from batch to batch. Note also that there are some spikes in the predicted activity. This means that some process measurements are noisy.

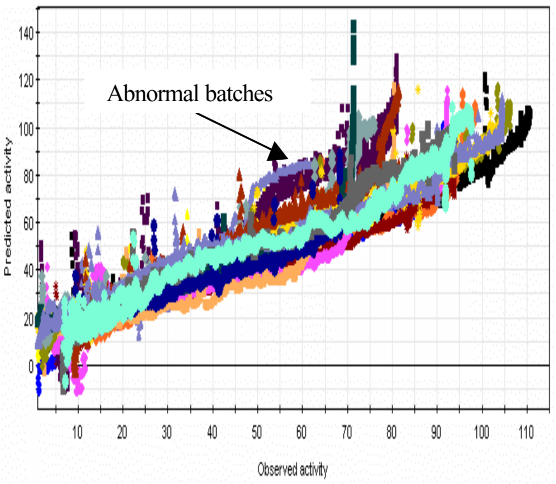

Fig. 4. Predicted activity versus observed activity for the baseline batches in the production phase.

Model disagreement is especially large for some of the batches at the end of production. For two batches, the predicted activity is higher than the observed activity for the entire production phase. This is a clear indication of abnormal batches.

3.2 Identification of bad batches

To understand the consistency of all the variables in the growth phase, a number of diagnostics tools can be used. One of them is the plot of Y-block latent variable versus X-block latent variable. If nonlinear relationships are observed for all batches, then it will be more appropriate to apply non-linear PLS (Qin and McAvoy, 1992). However, if linear relationships are observed for most batches, but nonlinear relationships are observed for others, this indicates the presence of bad batches.

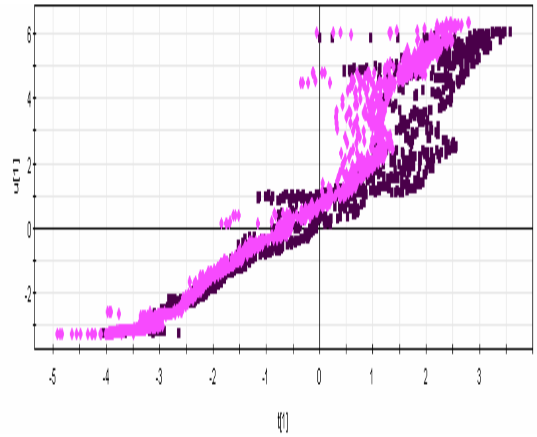

Fig. 5. Y-block latent variable versus X-block latent variable for two abnormal batches

As shown in Figure 5, nonlinear relationship is observed for two batches in the growth phase. With the use of the contribution charts (Miller and Swanson, 1998), many abnormal variables were revealed. As an example, one of these variables is plotted against OD for all batches in Figure 6. Three batches have lower values in this variable than the rest of the batches.

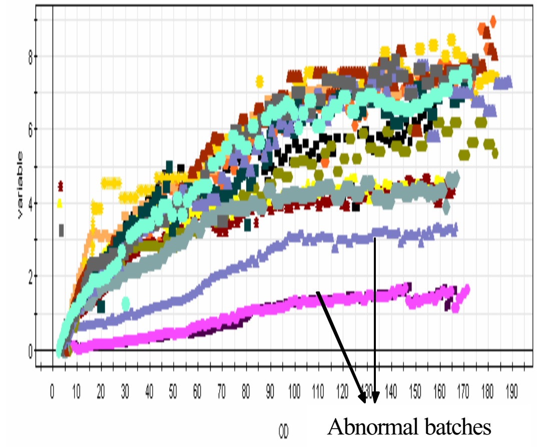

Fig. 6. Abnormal variable versus OD in the growth phase for all batches.

Diagnostic tools were also used to identify bad batches in the production phase. Two of the abnormal batches in the growth phase were also identified in the production phase. Operators confirmed that they encountered operating problems during these 3 batches. After removal of the inconsistent batches, the model fit for both phases improved. An additional result supporting the removal was that now the Y-block latent variable versus X-block latent variable is linear for the rest of 15 batches.

3.3 Root cause determination

Root cause analysis was based on the assumption that the top variables in the VIP plot explained most of the batch-to-batch variability. Using this information and fermentation knowledge, a new procedure with several corrective actions was applied to the process. Experiments were performed to validate improvements in the key parameter. It is shown in Figure 7 that variability in this key parameter becomes significantly smaller after the improvement (standard deviation decreases from to 0.328 to 0.0732).

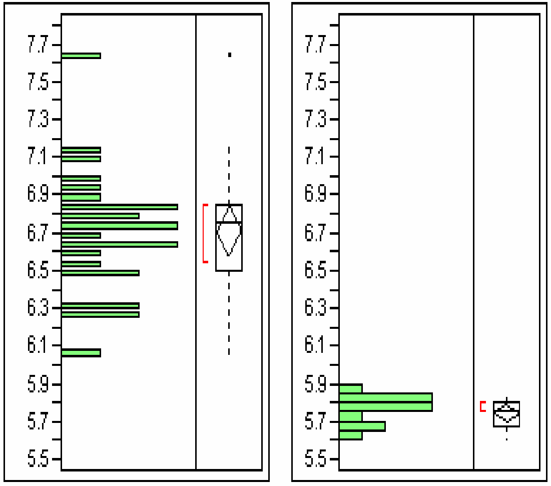

Fig. 7. Variability of the key variable before (left) and after (right) the implementation of new procedure.

3.4 Consistency for the batches under new procedure

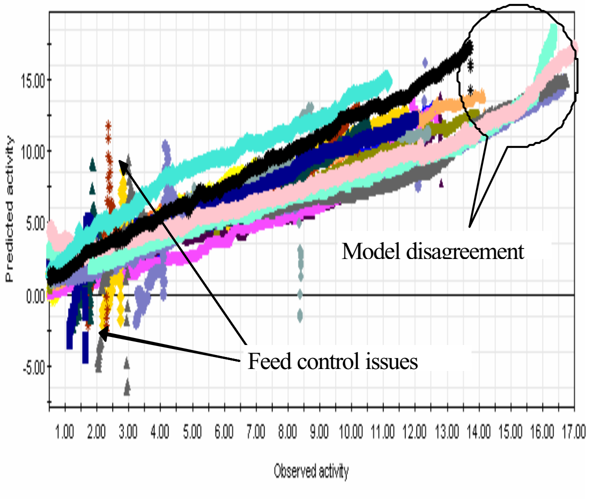

Fig. 8. Predicted activity versus observed activity for the new batches in the production phase.

Twenty batches were collected in order to validate the improvements. Of the 20 batches, seven batches are golden (no main problems were encountered and process data correlate well with the quality variables). One batch is known to be bad. The other 12 batches are questionable in a sense that some difficulties were encountered during the batch runs.

PLS models were applied here to identify the consistency of the batches. Improvement in the growth phase was validated. In the production phase the improvement is less clear on the first glance of Figure 8, which plots the observed and predicted activity ($R^2X = 0.635$ and $R^2Y = 0.866$ for a PLS model with 4 components). Six batches encountered feed control issues for two hours during the production phase. This is illustrated in Figure 8.

Figure 8 also illustrates another effect that is typical in real industrial applications. As it is shown, there are large model disagreements for seven batches. Without careful examination of the data and detailed discussion with the process engineers, one may incorrectly conclude that improvement has not taken place. However, a closer examination of the data shows that activity measurements for these seven batches are biased high. Large model disagreement means that a consistent model cannot be obtained to relate process measurements to the activity accurately for all 20 batches. This does not necessary mean that process measurements are inconsistent.

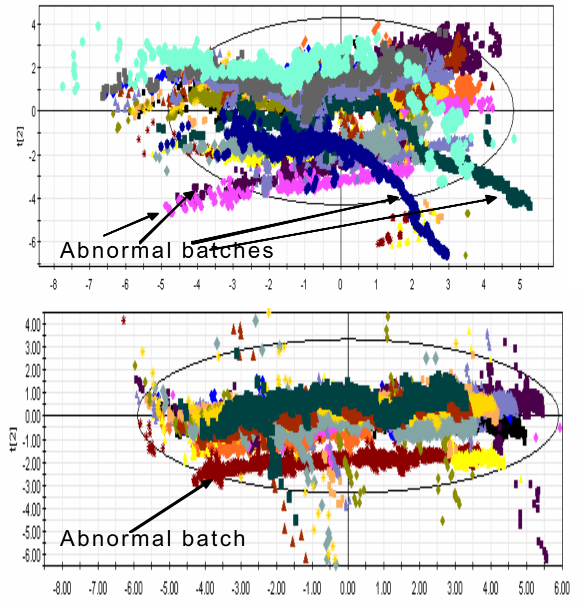

Fig. 9. Score plot consistency comparison between the baseline batches (top) and the new improved batches (bottom).

To understand the root cause, it is desired to examine the consistency of the process measurements using the score plot (see Figure 9). Each color in the plot represents a batch trajectory. Consistency in the batch (process data) is reflected in the consistency of the batch trajectories. In other words, if the process data are consistent among all batches, their batch trajectories will overlap. Of the 18 baseline batches (top of Figure 9), four batches (22%) clearly have different trajectories. This indicates that some process measurements are inconsistent from the rest of the batches. For the 20 new improved batches (bottom of Figure 9), only one batch (5%) has a different trajectory (this batch is known to be bad). The spikes in the score plot are results of known feed control issues. Process data were more consistent for the new batches, which validated the improvements that were implemented.

From the score plot, it is clear that the process measurements are consistent for the new batches. Because the same fermentation procedure was applied for these seven batches, activity measurements were expected to be similar. This posted a question of the activity measurements for these batches.

After discussion with the process engineers, it became apparent that the deviation in activity was due to a change of analyst. The measurements were precise but not accurate. Because of this, it is not meaningful to interpret the model fit and to compare the results to the baseline batches. Model fit for the new batches would improve if the hardware problems did not occur and activity measurements were not biased. Although the model fit cannot be compared directly, an examination between Figure 4 and Figure 8 shows that the activity prediction for the new batches is less noisy (spikes in the predicted activity disappear in Figure 8). This is an indication of improvement in the measurement system.

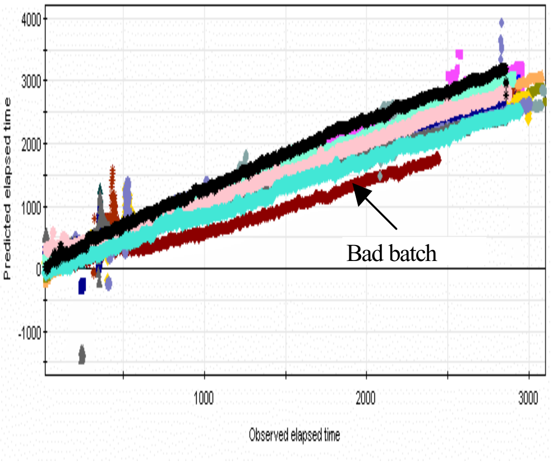

Fig. 10. Predicted elapsed time versus observed elapsed time for the new batches in the production phase.

To validate that the process data are indeed more consistent for the new batches than the baseline batches and that activity measurement are biased for some batches, additional PLS models were built with elapsed time, rather than activity, as the maturity variable. For these models, we are exploring the correlation between the process data and elapsed time for all batches. As shown in Figure 10, the model fit for the new batches (R^2X =0.829; R^2Y = 0.957 for a PLS model with 9 components) is better than the model fit in the baseline batches (R^2X =0.681; R^2Y = 0.955 for the model with 6 components) elapsed time as the maturity variable). This indicates that the process data are correlated with elapsed time consistently for most of the new batches.

For passing quality testing for a batch, six specifications have to be met. There are 18 baseline batches, which translates to 108 opportunities for failure. The baseline performance is 22 defects out of 108 opportunities (20.4%). For the 20 new batches, there are 4 defects out of 120 opportunities (3.3%) Therefore, quality improvement is clearly demonstrated.

4. ON-LINE MONITORING

To monitor performance during a batch run, it is effective to use on-line multivariate control charts. To illustrate this concept, golden batch data were used to build multi-way PLS models. The models capture normal variability of the batches that are acceptable to the users. Unlike Figure 9, the 7 golden batches overlap in the score plot (see Figure 11), which confirms that the process data are consistent for these batches.

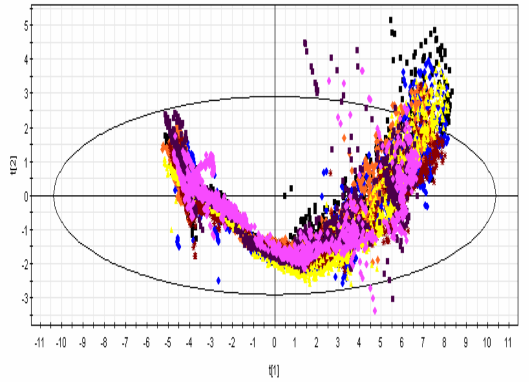

Fig. 11. Score plot for the 7 golden batches in the growth phase.

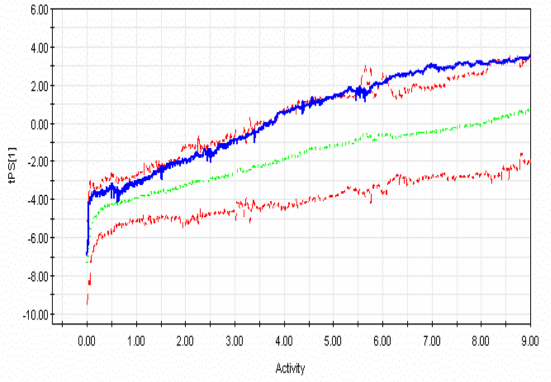

Fig. 12. Detection of an abnormal batch (solid line) using the first latent variable on-line. Dash lines and dotted line represent the critical limits and the average of the 7 golden batches, respectively.

After the models are built off-line, on-line monitoring takes place. Figure 12 shows the first latent variable for monitoring abnormal behavior during the growth phase of one bad batch. As process data are collected, on-line predictions of the latent variables and the quality variables are computed instantly. The critical limits (dashed lines) are defined based on the 3 standard deviation limit of the golden batches. The statistic was above the limit in the beginning for this batch, and then it went below the limit, and finally the statistic gradually increased above the limit again. Because no corrective actions were done during the batch run, the batch remained out of control. To identify abnormality during this run, contribution charts were used.

One of the variables that was identified in the contribution chart is plotted in Figure 13, which clearly shows that variable 10 (solid line) is above the average (dotted line) of the 7 golden batches. If online monitoring model was implemented during the batch run, problems could have been identified promptly, which could in turn avoid the production of out-of-specification product for this batch.

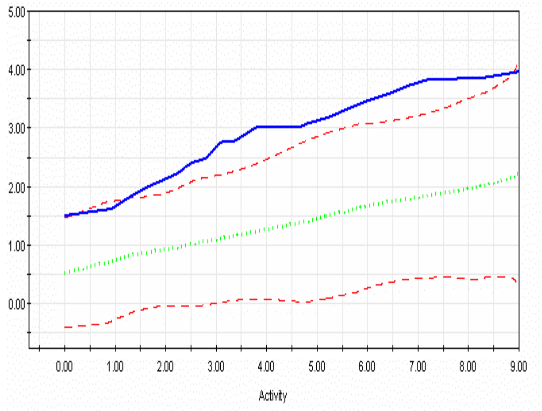

Fig. 13. Variable 10 (solid line) in an abnormal batch run. Dash lines and dotted line represent the critical limits and the average of the 7 golden batches, respectively.

5. CONCLUSIONS

Multi-way PLS was applied to 18 baseline batches of an industrial fermentation process at the San Diego biotech facility of The Dow Chemical Company. Three batches were identified as abnormal. The variability between the remaining batches was more significant in the production phase than the growth phase. The top variable in the VIP plots explained most of the variability. A new procedure was carried out to control this key parameter and twenty new batches were made with the new procedure. Multiway PLS analysis showed that the new batches were more consistent in both phases as evidenced in the improvement in the model fit (R^2X = 0.829 for the new batches versus R^2X = 0.681 for the baseline batches) and in the percent of out-of-specification (3.3% for the new batches versus 20.4% for the baseline batches).

Seven golden batches were used to build a PLS model to capture normal operation conditions. Bad batches were detected promptly and abnormal variables were determined accurately.

The use of multivariate analysis plays an important role at The Dow Chemical Company. Value has been realized in many applications including batch quality improvement (this paper), scale-up of new batch agrochemicals (Schnelle and Armstrong, 2003), variable selection for multivariate calibration (Leardi *et al*., 2002), and soft sensor development (Kordon *et al*., 2002). Implementation of on-line batch monitoring in the future is expected to bring further value.

ACKNOWLEDGMENT

Karl Schnelle and Mary Beth Seasholtz are thanked for their inputs on the manuscript.

REFERENCES

Beebe, K. R., R. J. Pell and M. B. Seasholtz (1998). *Chemometrics: A Practical Guide*, John Wiley & Sons.

Chiang, L. H., E. L. Russell and R. D. Braatz (2001). *Fault Detection and Diagnosis in Industrial Systems*, Springer-Verlag.

Eriksson, L., E. Johansson, N. Kettaneh-Wold and S. Wold (2001). *Multi- and Megavariate Data Analysis,* Umetrics Academy.

Kordon, A. K., G. F. Smits, E. Jordaan and E. Rightor (2002). Robust soft sensors based on integration of genetic programming, analytical neural networks, and support vector machines, *Proceedings of WCCI*, 896 - 901.

Leardi, R., M. B. Seasholtz and R. J. Pell (2002). Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. *Anal. Chim. Acta*, **461**, 189-200.

Miller, P. and R. E. Swanson (1998). Contribution plots: a missing link in multivariate quality control. *Appl. Math. and Comp. Sci.,* **8,** 775-792.

Qin, S. J. and T. J. McAvoy (1992). Nonlinear PLS modeling using neural networks, *Comput. & Chem. Engr.,* **16**, 379-391.

Schnelle, K. and K. Armstrong (2003). Statistical analysis of large pilot plant datasets, *Proceedings Foundations of Computer-Aided Process Operations (FOCAPO)*, 459-462.

Umetrics, Inc. (2003). SIMCA-P+, version 10, Multivariate process modeling software, www.umetrics.com.

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

Zhang, H. and B. Lennox (2004). Integrated condition monitoring and control of fed-batch fermentation processes. *J. of Process Control,* **14,** 41-50.