STATISTICAL ANALYSIS OF LARGE PILOT PLANT DATASETS

Karl D. Schnelle* and Katherine Armstrong
Dow AgroSciences LLC
Indianapolis, IN 46268 USA

Abstract

Dow AgroSciences has supported two pilot plant facilities for development and scale-up of new batch agrochemicals for several years. One of the pilot plant facilities consistently obtained higher yields and lower run-to-run variability. An improvement project using recently developed data mining techniques and Six Sigma methodology was initiated to resolve the operating differences between the two plants. The project goal was to decrease the coefficient of variation (CV) of the pilot plant B to 6% and improve the mean yields to within 95-100% of pilot plant A.

Therefore, pilot plant datasets were acquired and analyzed using two multivariate statistical techniques, Principal Component Analysis (PCA) and Partial Least Squares (PLS). These techniques identified variability both between batches and across time for each batch. The most significant input variables for overall variability were three raw materials and one operating condition. Based on this information and operator experience, experiments were run at B resulting in lowering one feed rate and raising another. Because of these improvements, Dow AgroSciences discontinued the use of plant A, which resulted in significant savings per year. Even greater savings has been leveraged into manufacturing from less raw material use.

Keywords

Principal Component Analysis, Partial Least Squares, coefficient of variation, design of experiments

Introduction

Originally, Dow AgroSciences negotiated studies through a contract with pilot plant A that capitalized on their extensive expertise. In addition, a pilot plant B was built to develop in-house expertise. The mission of this facility was to validate yield improvements and optimize production conditions.

Pilot plant A consistently obtained higher yields and lower run-to-run variability than B pilot plant. An improvement project using Six Sigma methodology (Wheeler, 2002) was initiated to resolve the operating differences between the two plants. The project goal was to decrease the CV of pilot plant B to 6% and improve the mean yield for new products to 95-100% of pilot plant A. This research paper focuses on the process data analysis phase of the project, which helped to determine the significant variables affecting the CV and mean yield.

Over a six-month period, 1500 records were obtained over time for both control and production lots at A and B. The data were analyzed by using Principal Component Analysis (PCA) and Partial Least Squares (PLS). These multivariate statistical techniques accounted for variability not only between batches but also across time for each batch.

Organization of Pilot Plant Data

Batch run data and laboratory sample analysis data were collected for both A and B pilot plants during year

* To whom all correspondence should be addressed. *E-mail address:* kschnelle@dow.com

2000. Both control lots and normal production lots were collected and organized into one data spreadsheet for the same product. Table 1 illustrates the types of lots.

Table 1. Number of lots used in the analysis

Plant	Control Lots	Production Lots
A	8	42
B	25	46

The different types of variables acquired are listed in Table 2. Several batch run variables, in Table 3, were identified by the Six Sigma team as possibly significant to yield. These were reformatted from ASCII or spreadsheet hourly batch sheets from each plant into a single spreadsheet matrix for 12 time steps 1 to 262. Because of the difference in reactor sizes between the two plants, feed rates were ratioed by batch size.

Table 2. Data matrix variable types

Type	Number
Batch run variables	16
Input sample variables	6
Output sample variables	11
Final batch size	1
Total	34

Table 3. Batch run variables

Type	Number
Operating characteristics	9
Amount of raw material	4
Raw material feed rate	3

The time steps that were acquired for each batch run variable were selected to correspond to the times when samples were pulled and assayed from the process; time step 1, 22, 46, 70, 94, 118, 142, 166, 190, 214, 238, and 262. Therefore, input sample (batch characteristics) and output sample variables (yield, amounts of intermediates, impurities, and product) were added onto the batch run data by lot for each of the 12 time points to form the final data matrix. Final batch size in the reactor was also recorded.

Analysis Methods

Motivation for Multivariate Methods

Once the pilot plant data were acquired and organized into the matrix, the task of analysis began. The goal was to make full use of all the data and to limit the misinterpretation of that data. Several different approaches to statistical analysis could be performed on the data. However, with large datasets (in this case, 121 total lots x 12 time steps x 33 total variables), a multivariate statistical analysis using PCA/PLS was much more efficient to conduct than a traditional linear regression or correlation analysis. Thus, one avoided the trap of "data overload" with univariate methods. The PCA/PLS method could model both multiple inputs as well as multiple outputs. Some of the advantages to this multivariate analysis were that it could handle large dimensionality, collinearity, noise, outliers, and missing data. Thus, this approach was used to gain as much insight as possible into the process being studied.

PCA reduced the dimensionality of the large data matrix by explaining variance with new Principal Components, PCs. The first factor, or PC, was obtained by finding the linear combination of variables explaining the greatest amount of variability in the data. The second principal component was obtained by finding another linear combination of patterns that is at “right angles” (i.e. orthogonal and uncorrelated with) to the first principal component. Each succeeding PC was similarly obtained. Thus, the information in the original large dataset was represented by a much smaller number of new PCs. There would never be more PCs than there are variables in the data. PLS reduced the dimensionality of an input matrix X and an output matrix Y; this method found PCs that both model the variance in X and correlate to Y. Thus, Y could be predicted from X. Kourti, Nomikos, and MacGregor (1995) were some of the first researchers to apply this method to batch process data; their major research is reviewed in Westerhius, Kourti, and MacGregor (1999).

Multi-way PLS was a special two-level method that could reduce both variables and time steps into new PCs. The higher level was a PLS run on all batch variables over time as the X matrix, with time as the only variable in the Y matrix. This created a set of new PCs that account for both batch and time variability. Then, a lower level PCA was executed to determine variables that were significant to the variability (Wold, *et al*, 1998). Also, final yield and amount of product could be used in a Y matrix in the lower level, with the same PCs as the X matrix. Then, a lower level PLS was run to determine variables that affect the final lot yield or amount of product (Kourti, Nomikos, and MacGregor, 1995).

Application to Pilot Plant Data

Because the goal of this analysis was to determine which variables affect the yield, the 16 batch run variables were combined with 8 of the sample variables that describe process conditions to form the input matrix X. With final batch size, the input matrix had 25 total variables for the higher level PLS. The output matrix, Y, for the lower level PLS was comprised of four significant sample values related to yield, amount of intermediates remaining, and final product.

Figure 1 shows the batch run data as block A being combined with sample data B. Multiway PLS was

performed on this combined dataset with time step as the only Y variable. The results of the PLS were a set of PCs for each lot that represent the variability in the original input variables over time. Then the PCs were combined with final batch size, yield, and amount of product, block C, so that the lower level PLS may be performed to determine the significant variables that affect yield. SIMCA-P™ 8.1 (Umetrics, 2001) was the software tool used in this analysis that incorporates both PCA and PLS for analysis of batch chemical processes.

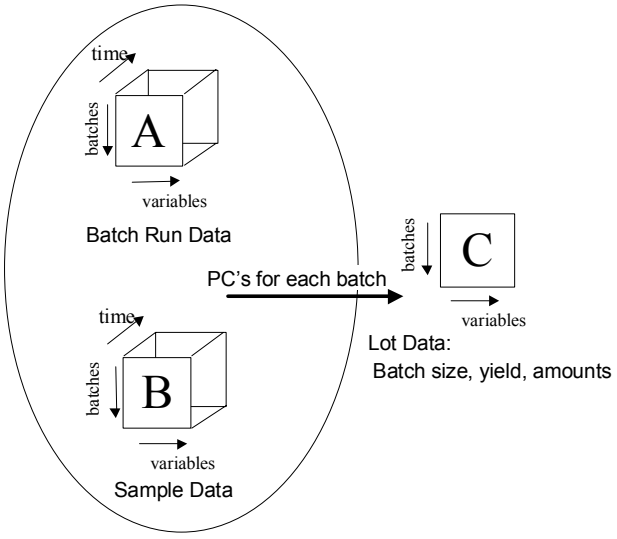

Figure 1. Types of datasets

The overall flow of information from data acquisition to analysis is shown in Figure 2. Because control lot data was available that would represent consistent operation inside each plant better than all lots in general, the analysis was first performed on just the control lots. Any difference in normal operation between A and B should be more apparent in this analysis than in the overall lot analysis.

Summary of Results

Control Lots

Significant differences between the 33 control runs at B and A were found using multiway PLS. Four principal components (at each time step) were found to predict time with R^2=0.78. (This value was an indicator that the first four PCs model 78% of the variability in the X matrix.) Batch size was most different; this result was obvious because A utilizes larger tanks. Then, five variables were found to be different at different times during the runs: three operating conditions at the beginning and middle, then two intermediates at the end of the runs.

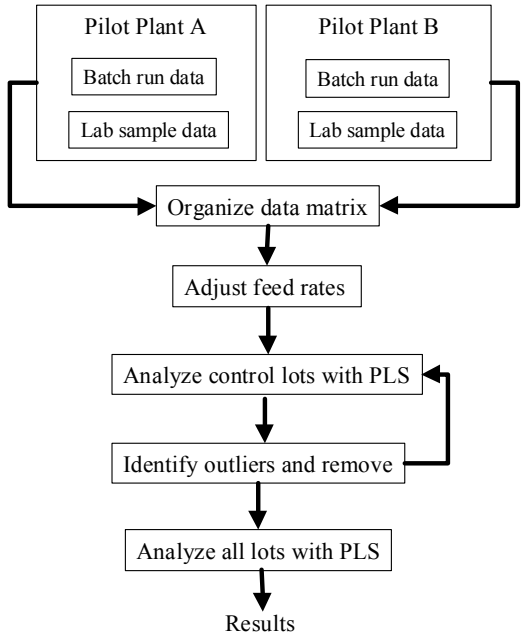

Figure 2. Information Flow

Then, the B control runs were analyzed separately to study the major sources of variability within B. Again, 4 PCs were found with almost the same fit: R^2=0.76. Including all the variables, the four highest contributors from the lower level PCA were not controllable in the pilot plant, but one of the raw material feed rates was the most significant variable that could be controlled. Plant A did not conduct enough control runs for this product to find any significant variables.

All Lots

The higher level PLS was executed for 121 lots that resulted in 1560 observations (all lots at each time step). Four PCs (at each time step) were found to predict time with R^2=0.79. From the lower level PLS, two PCs were found with R^2=0.69. The most significant input variables for predicting final product amounts were three raw material feed rates and two uncontrolled operating conditions. Q^2 was the measure of how well the inputs predict the outputs in the lower level PLS. For this model, Q^2=0.49; any model with a Q^2 of 0.50 or higher can be considered good.

Detailed Results

Analysis of Control Data

Four PCs were required to form a significant Multiway PLS model for the control lots. The second PC showed A and B lots in two different clusters. The separation was easily observed. By diagnosing the cause

of the two clusters, five variables (those with the largest difference in weights on the second PC) were identified that cause the difference over time.

Then, A and B control runs were analyzed separately to study the major sources of variability within each. No clusters in the PCs could be seen for either plant individually. Because of the low number of control lots at A, a good PLS model could not be found. For B, the lower level PCA determined that the top variables that contribute to variability are four variables from the batch run data. The most significant variable was plotted over time to determine if any outliers affected the results. The value for one batch at the end of the run did not increase with the other lots; it was removed and the analysis repeated. However, the same significant variables were found, indicating that this outlier did not affect the results.

The analysis for B included all variables, so the output variables were excluded and the analysis rerun to determine if any adjustable inputs would be significant. Again the top four variables, from the lower level PCA, were the same four batch run variables. The most significant was plotted over time again. Large dips at certain times indicated bad probe readings due to possible poor calibration. So this variable was deleted, and one lot deleted due to processing problems. The same top four significant variables were found as before. Therefore, no new information was gathered by deleting suspected outliers.

The four significant variables were not directly controllable in the plant, but the next significant variable that can be controlled was a feed rate. Potential outlier lots (after time 165) were investigated. Three lots had invalid feed rates and were deleted. The other lots were left in, and the analysis was re-run. No changes in the significant variables were detected.

Analysis of Data for All Lots

The higher level PLS produced a good model, but the 121 lots did not cluster together for any of the PCs, as they did for the control lots, *i.e.*, the PCs over time were very consistent. Significant variables from the lower level PLS model were found.

Instead of plotting PCs over time for each lot, the first PC for the X matrix, from the lower level PLS model, was plotted versus the first PC for the Y matrix. Now clusters could be seen where each point represents a lot. Five B lots were clustered closer to the A batches. By comparing the difference in significant variables between those five and the A lots, three operating variables were found to cause the different clusters. These five B lots did in fact use a different recipe than the other B lots; that recipe was closer to A's.

Because the other lots clustered according to location, two separate models were built by location. However, no additional information was gained.

Experimental Runs at the Pilot Plant

From the Control Lot analysis, three variables were identified as potential causes for the lower yields and increased variability in pilot plant B. A design of experiments (DOE) was carried out in B to look at these three factors. Increasing one feed rate consistently produced higher yields. This rate, along with an operating condition, also affected variability. The other feed rate did not affect yield and was reduced.

Conclusions

Multivariate analysis proved to be a powerful statistical tool to evaluate complex batch data generated by the two pilot plants. The results of the analysis allowed the team to focus on the critical variables. Based on operator experience, designed experiments were run at pilot plant B which found improved yields at B to within 8% of pilot plant A. These results were validated over multiple batches, and the CV decreased to 4%.

The Six Sigma process provided structure to address the yield discrepancies between the two pilot plants. Although statistical tools are used routinely at the pilot plants, the novel multivariate modeling tools helped focus on the critical variables and confirm that the scientists were on the right track toward improving operations.

Acknowledgements

The authors wish to thank Tony Tietz and Gary Kleman, who provided organized data and thoughtful feedback, and Kent Steele, who initiated the project and provided statistical insight. We would also like to thank Dawn Booms, Kathy Koch, Dave Harp, and Dave Osentoski for generation of batch data and Dave Sieman for analytical evaluations.

References

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

Wheeler, J. M. (2002). Getting Started: Six Sigma Control of Chemical Operations, *Chem. Eng. Prog*, June, 76-81.

Westerhius, J. A., T. Kourti, and J. F. MacGregor (1999). Comparing Alternate Approaches for Multivariate Statistical Analysis of Batch Process Data, *J. Chemometrics*, **13**, 397-413.

Wold, S., N. Kettaneh, H. Freden, and A. Holmberg (1998). Modeling and Diagnostics of Batch Processes and Analogous Kinetic Experiments, *Chemometrics and Intell. Lab. Systems*, **44**, 331-340.

Umetrics, Inc (2001). SIMCA-P V8.1 software for multivariate modeling and analysis, www.umetrics.com.