Combining Process and Spectroscopic Data to Improve Batch Modeling

Jon Gabrielsson, Hans Jonsson, and Johan Trygg
Research Group for Chemometrics, Umeå University, SE-901 87 Umeå, Sweden

Christian Airiau, Bernd Schmidt, and Richard Escott
GlaxoSmithKline, Tonbridge, Kent, U.K.

DOI 10.1002/aic.10932
Published online June 29, 2006 in Wiley InterScience (www.interscience.wiley.com).

Pharmaceutical production is at present characterized by static processes where quality is guaranteed by controlling the purity of the final product. Achieving better control throughout the process, as a means for improving product quality, is one of the objectives of the PAT initiative by the FDA. A data set consisting of 11 batches characterized by UV spectroscopy together with process data was used in this study. Design of experiments was used to introduce controlled process variation in test batches. The objective was to investigate possible advantages of MSPC using a combined data set, compared to separate models of the respective data sets. Individual models for the separate data sets show that they contain complementary information. A major advantage of combining spectroscopic and process data is that deviations that would go unnoticed using just an individual model can be detected and interpreted. All process manipulations were detected by the combined data set model. Implementation of these methods to batch processes in primary and secondary pharmaceutical production is feasible. An enhanced understanding of the process together with control tools should lead to a well-understood process and, ultimately, real time release. © 2006 American Institute of Chemical Engineers
AIChE J, 52: 3164–3172, 2006
Keywords: PAT, design of experiments, pharmaceutical batch process, multi-block data, process monitoring

Introduction

The process analytical technology (PAT) initiative by the Food and Drug Administration (FDA) has highlighted the importance of evaluating not only the final product but the whole production process.[1] At present, multivariate data are routinely collected from process analyzers in the pharmaceutical industry. What to do with the data and how to best use them in order to improve production technology now becomes an important challenge.

The physical state of a process is derived from measured variables, for example, pH and temperature. Monitoring of these process variables using traditional approaches only makes use of one variable at a time. Statistical process control (SPC) tools such as Shewhart or Cumulative sum (CuSum) plots can be used to display the information to the process operator. The state of a process is often determined by several factors, and possible interactions among the measured variables cannot be detected with a univariate approach.[2] Significant advancements have been made within the last two decades, and multivariate statistical process control (MSPC) is now a well-established methodology for analysis of batch data.[3,4] The advantages of using the multivariate methods over the more traditional SPC methods also include the ability to

Correspondence concerning this article should be addressed to J. Trygg at johan.trygg@chem.umu.se.

handle missing data and collinearities among measured variables.[5] Non-invasive spectroscopic methods, for example, ultraviolet (UV) and near infrared (NIR), can give additional information to the process variables regarding the chemistry of the process.

The use of design of experiments (DoE) and multivariate methods together with spectroscopic process analyzers has proven to be successful.[6,7] Today process information representing both physical and chemical information is available. The challenge is to merge the different types of data. In this study, the combination of process variables and spectroscopic profiles will be evaluated. Combining process and spectroscopic variables into one MSPC model will enhance the possibilities of fault detection and outlier detection, and provide easier diagnosis of possible process deviations. The result is an improved understanding of the process, and DoE derives the important factors that control it, which is in line with the PAT initiative by the FDA.

Some process faults are invariably caused by physical attributes such as temperature and pressure, while others are caused by chemical attributes such as poor starting material or catalyst, or by a combination of both chemical and physical factors. The approach discussed here can provide a better characterization of the process, thus reducing the frequency of false alarms to the process operator.

The objective of this article is to investigate the advantages of an MSPC model based on a combination of process and spectroscopic variables to improve the understanding of the process. DoE is used to introduce controlled process variation in the test batches. A variety of controlled process manipulations is used to evaluate the model.

Simultaneous monitoring of both the chemistry and the physics of the process also leads to improved knowledge regarding critical process variables, hence increasing the possibilities of active process control.

Methods

Short descriptions of methods used in this study are given below. For a thorough account of the theoretical aspects of the methods used, the reader is referred to the provided references.

Design of experiments

DoE is a framework for systematic evaluation of complicated systems that are influenced by multiple factors, for example, pH, temperature, and amount of catalyst. The objective

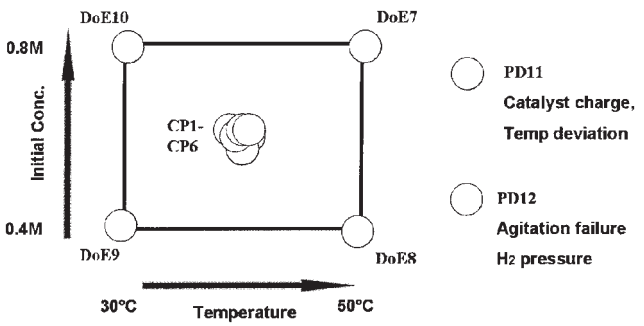

Figure 1. Illustration of the 2^2 full factorial design that was implemented in the study.

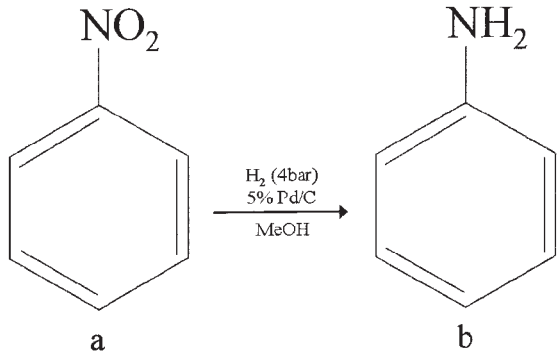

Figure 2. Reaction scheme for the conversion of Nitrobenzene (a) to Aniline (b).

of DoE is to effectively plan and conduct experiments so that the experimental domain is systematically investigated with as few experiments as possible. Experimental factors and the range in which they are varied define the experimental domain, the area that is investigated. The response variables, for example, yield and side product, are measured results of the performed experiments.

The factors can be either quantitative or qualitative. A quantitative variable is a continuous variable that can take any number between predefined levels in the design. A factor that may only be varied at distinct levels such as present/not present, on/off, or Excipient A or B or C is termed a qualitative factor.

In a full factorial experimental design, all k factors are changed simultaneously, but in an orderly fashion. This makes it possible to examine main effects as well as interaction effects. The geometrical representation of the experimental design in the case of two factors is a square; with three factors it is a cube; four factors make up a hypercube, and so on (Figure 1). If k variables are investigated at two levels, the number of experiments in the full factorial design is given by the expression 2^k. There are several textbooks and articles regarding theoretical aspects of experimental design.[8-10]

Multivariate methods

A graphical illustration of an $N \times K$ data matrix is a swarm of N points, the observations, in a K-dimensional coordinate system. In practice, variables are highly correlated, especially in the case of spectral variables, where a high absorbance at one wavelength is usually accompanied by similar absorbance values at neighboring wavelengths. This collinearity reduces the K-dimensional space to a much smaller subspace.

Graphically, Principal Component Analysis (PCA) corresponds to a least squares fitting of a straight line ($A = 1$), a plane ($A = 2$), or an A-dimensional hyper plane ($A > 3$) to the data in the K-dimensional variable space.[11] The observations are projected onto a subspace of lower dimension, and the variation of the observations is summarized in the $\mathbf{T}$ ($N \times A$) matrix, which consists of a score vector $\mathbf{t_a}$ for each component. How the original variables influence the principal component is summarized in the $\mathbf{P}$ ($K \times A$) matrix. The loading vector $\mathbf{p_a}$ describes how the variables are correlated and explains the observed trends and groupings in the scores. The difference between the original coordinates and the projections are termed

Table 1. Factor Settings for the 2^2 Full Factorial Experimental Design for the Conversion of Nitrobenzene to Aniline

Batch Name (abbreviation)	Reaction Temperature (°C)	Initial Conc. (mol/L)	Duration (min)
Center point 1 (CP1)	40	0.6	45.7
Center point 2 (CP2)	40	0.6	39.1
Center point 3 (CP3)	40	0.6	37.8
Center point 4 (CP4)	40	0.6	37.9
Center point 5 (CP5)	40	0.6	38.9
Center point 6 (CP6)	40	0.6	37.3
Experiment 7 (DoE7)	50	0.8	24.1
Experiment 8 (DoE8)	50	0.4	—
Experiment 9 (DoE9)	30	0.4	66.1
Experiment 10 (DoE10)	30	0.8	45.4
Process deviation 11 (PD11)	40	0.6	47.6
Process deviation 12 (PD12)	40	0.6	33.9

Included also are six center points and two batches with introduced process deviations (explained in detail in Table 2). The resulting duration of each batch is given in min.

residuals. The residual matrix **E** contains the part of the data matrix that is not explained by the model. From this matrix, the distance to model in **X** (DModX) can be calculated, which shows how well observations fit into the model.

Score values from two principal components, for example, $\mathbf{t_1}$ and $\mathbf{t_2}$, form a two-dimensional model of the data. One could say the $\mathbf{t_1}/\mathbf{t_2}$ score plot constitutes a window through which data can be viewed. This facilitates the detection of groupings, trends, and outliers (deviating observations) in data. Many descriptions of PCA can be found in the literature.[11-13]

Partial least squares projections to Latent Structures (PLS) is a regression extension of PCA. The objective is to find the latent structure in the **X** matrix (the descriptive variables) and the **Y** matrix (the response variables) to maximize the covariance between the matrices. For each component a weight vector, **w**, containing the contributions from the descriptive variables to the explanation of **Y** in that particular component is calculated. The corresponding weight vector for **Y** is termed **c**.

Data from batch processes include measured variables that vary over time and are measured at regular intervals for a number of batches. The batch trajectory can be estimated using Multi-way PCA or Multi-way PLS[14] based on, for example, process variables or digitized spectra. By applying basic statistics to the resulting latent variables, control charts can be constructed. The MSPC model can then be used for analysis of batches, diagnosis of deviating batches, and predicting the evolution of new batches.

The quality of a model can be quantified into a measure of explained variation, R^2, and Q^2, a measure of the model predictive ability by means of cross validation. The Q^2 value is calculated according to Eq. 1. The data are divided into seven cross validation groups consisting of entire observations.

$$Q^2 = \frac{SS_Y - PRESS}{SS_Y} = 1 - \frac{\Sigma(Y_{obs} - Y_{pred})^2}{\Sigma(Y_{obs} - Y_{mean})^2} \quad (1)$$

For further reading on PLS, please refer to refs. [15-18].

In practical situations, variation can exist in **X** that is not associated with **Y**, often exemplified as baseline and unknown impurities. This causes problems ranging from difficulties in interpretability, for example, models that contain additional components,[19] to lack of established relationships due to high noise levels.

Pre-processing and scaling of multivariate data

The overall aim of data pre-processing of spectroscopic profiles is to remove unwanted systematic variation, such as baseline shifts and scatter effects. Pre-processing methods commonly used for spectral data include differentiation and scatter corrections methods, for example, multiplicative signal correction (MSC) and standard normal variate transformation (SNV).[20-22]

Projection methods such as PLS are scale sensitive. The main reason for scaling data prior to analysis is to reduce the influence of noise. If scaling is applied in a careful manner, it facilitates the possibilities of obtaining reliable results in data analysis. The scaling of a variable or blocks of variables should reflect prior knowledge of the individual variables.[23]

Mean centering the respective columns is applied to remove the mean trajectory of the data. Additionally, unit variance scaling is applied to remove differences in range between variables. This involves a risk of scaling up noisy variables and, for spectroscopic data such as UV and NIR, mean centering only is often appropriate. A possible drawback of using mean centering is that pertinent information that might be hidden in the noise is lost. Process variables, for example, reactor temperature and gas feed, often have different ranges, that is, they differ more than a factor of 10, and scaling to unit variance or similar is thus appropriate.

Block scaling is an alternative for facilitating modeling of multiple source data when blocks have different numbers of variables or variance. Each block is scaled to equal the square root of the number variables in that block. Another alternative

Table 2. Process Disturbances Introduced for Batches PD11 and PD12

Batch Name (abbreviation)	Description of Process Deviation
Process deviation 11 (PD11)	10% less catalyst - cooling to 30°C at t = 7 min with jacket control, then (t = 15 min) with internal control - heating to 40°C at t = 20 min - ramping from 40°C to 31°C in 3°C steps: to 37°C (t = 25 min), to 34°C (t = 37 min), and to 31°C (t = 48 min)
Process deviation 12 (PD12)	stop stirrer at t = 7 min, restart at t = 8 min - decrease pressure from 400 kPa to 200 kPa from t = 13.3 min to t = 15.3 min and hold until t = 16.3 min - increase pressure to 400 kPa until t = 21.4 min - stirrer speed ramp to 750 rpm (t = 21.5 min to t = 24.5 min) - increase stirrer speed to 1500 rpm at t = 24.6 min

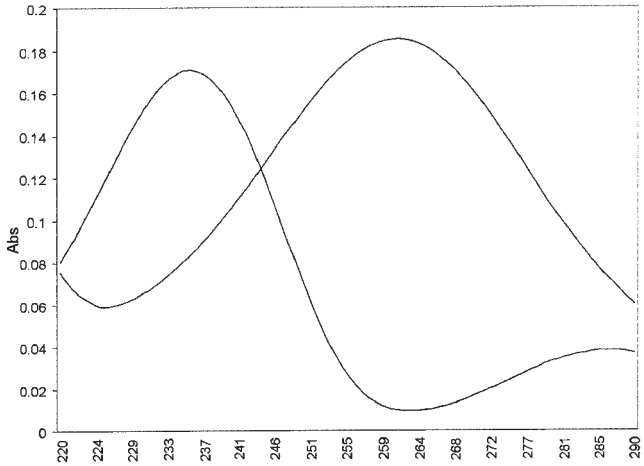

Figure 3. Raw UV spectra of the starting material Nitrobenzene (peak at 260nm) and the product Aniline (peak at 235nm).

is to perform a block scaling that equals the sum of squares in each block. This enables the use of mean centering only in one block, for example, spectroscopic data, and the use of both mean centering and unit variance scaling in another block, containing for example process parameters.

Experimental

The conversion of Nitrobenzene to Aniline[24] was carried out in a purpose built laboratory scale plant (Figure 2). The hydrogenation reaction was performed in a 1l-dish-based hastelloy vessel, fully baffled and running a pitched blade type impeller. Methanol (analytical grade, Fisher Scientific, Loughborough, UK) was used as solvent, and the catalyst was 5% Pd/C (87L paste, Johnson Matthey, UK). DoE was implemented in order to obtain controlled process variation (Figure 1). Six center points with the desired process settings were carried out. These center points will constitute the normal batches that will be used to build an MSPC model. The reaction times ranged from 37.3 to 45.7 min, with an average reaction time of 39.5 min. The initial concentration of the Nitrobenzene and the reactor temperature were varied according to a 2^2 full factorial statis-

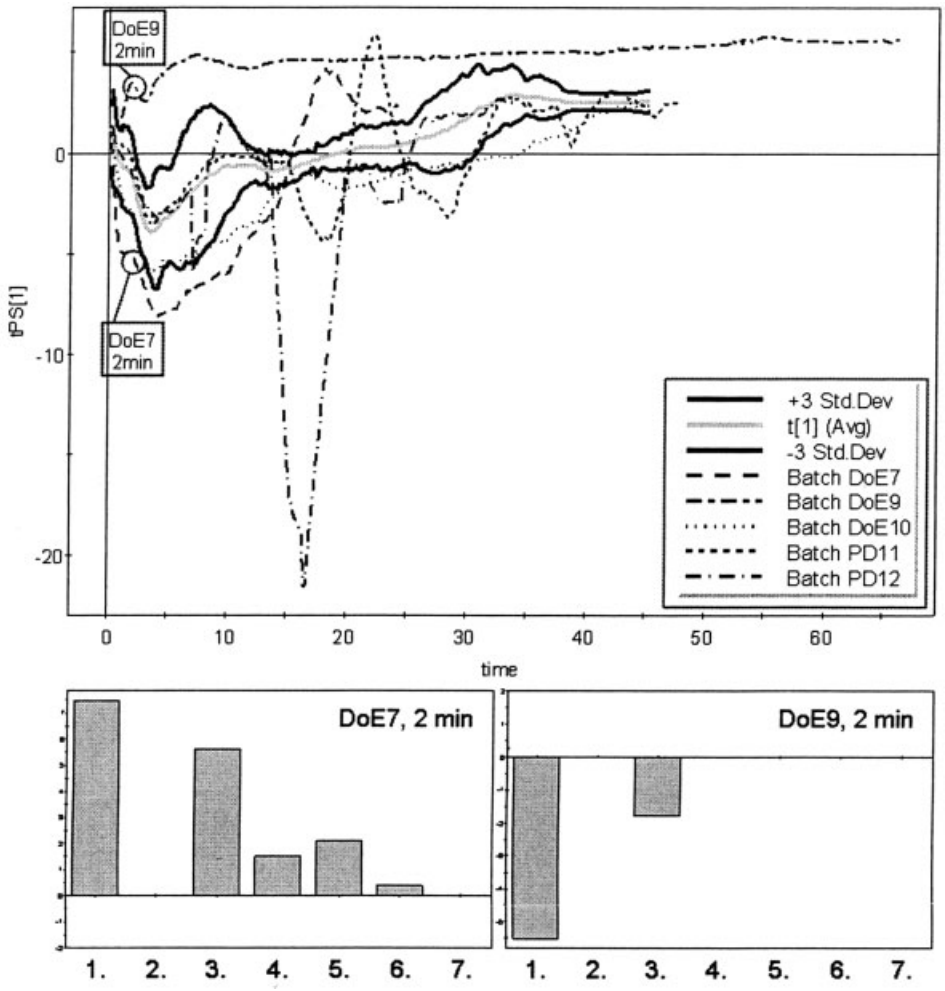

Figure 4. The boundaries in the batch control chart (above) correspond to average score value ± 3 standard deviations.

The score profiles of each batch in the test set are marked according to the scheme provided in the figure. Contribution plots from the first component of the PLS model of process data for DoE7 (lower left) and DoE9 (lower right) after 2 min. The process variables are: 1. reactor temperature, 2. reactor pressure, 3. gas feed, 4. jacket in temperature, 5. jacket out temperature, 6. flow rate of oil, 7. stirrer speed.

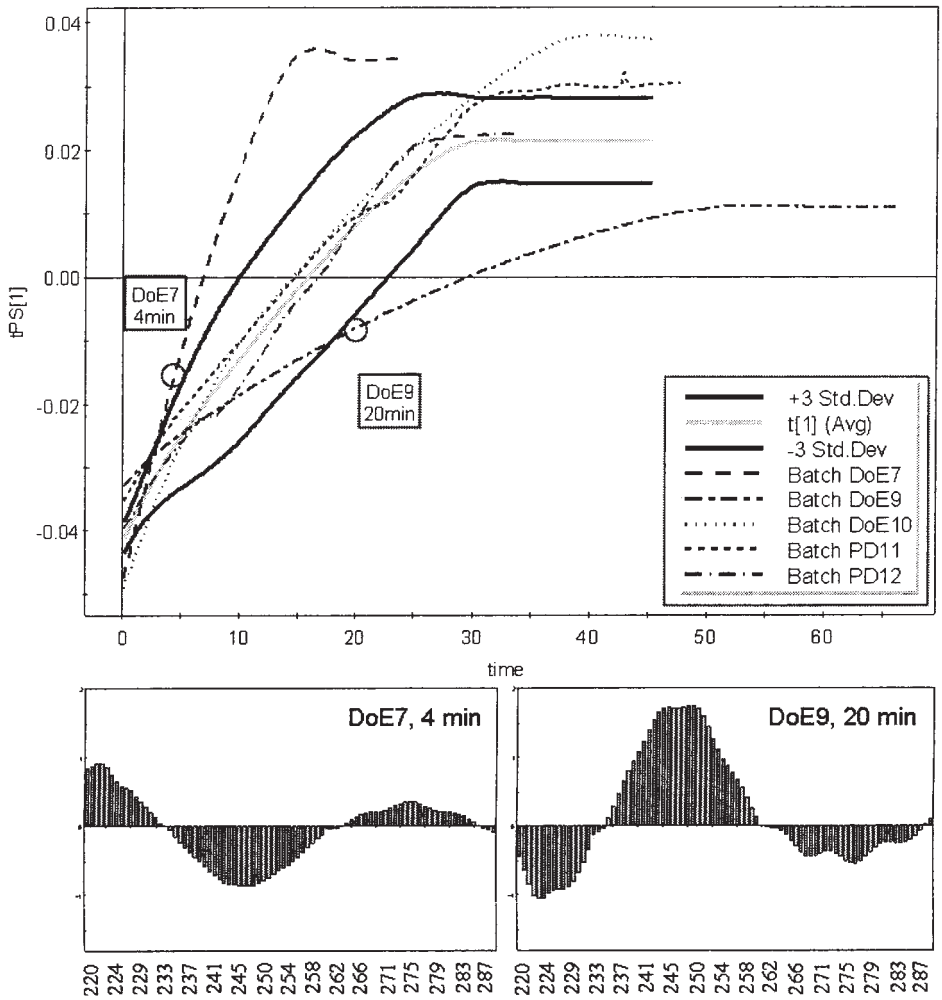

Figure 5. The boundaries in the batch control chart (above) correspond to average score value ± 3 standard deviations.

The score profiles of each batch in the test set are marked according to the scheme provided in the figure. Contribution plots from the first component of the PLS model of spectroscopic data for DoE7 after 4 min (lower left) and DoE9 after 20 min (lower right).

tical experimental design; see Table 1. In addition to the four experiments, two batches were varied both in terms of initial settings and reaction conditions (see Table 2) to simulate process deviations.

The reaction was triggered by the onset of H_2 gas flow, and process data were collected every 5 s. The process variables were reactor temperature (°C), reactor pressure (kPa), gas feed (L min^{-1}), jacket in temperature (°C), jacket out temperature (°C), flow rate of oil (L min^{-1}), and stirrer speed (rpm). For DoE9 there is missing data for jacket in temperature (°C), jacket out temperature (°C), and flow rate of oil.

An attenuated total reflectance (ATR) probe was installed on the reactor vessel, and the reaction was monitored by UV spectroscopy (Carl Zeiss diode-array MCS-500, Jena, Germany). Absorbance was measured in the UV region (220-400nm) every 10 s, and the spectra show a linear response for the investigated concentrations. Nitrobenzene absorbs around 260nm and Aniline around 235nm. Initial investigations found that wavelengths above 290nm were non-informative regarding the chemical reaction. For this reason, only the 220-290nm interval (84 variables) was used in the subsequent models (Figure 3).

Due to a UV spectrophotometer failure, no spectra were available for DoE8 and the batch was thus excluded from the study.

All PLS models were performed with Simca-P+ 11.0, and O2PLS calculations were performed with Matlab.[25,26]

Alignment, Pre-Processing, and Scaling

Alignment

Process and spectral data were recorded at different time resolutions; process data were logged every 5 s and a UV spectrum was recorded every 10 s. The data sets were merged with in-house built software using Matlab functions.[25] Alignment was performed with the nearest point method, and selections were based on the data/time stamp generated by each software, with both computers synchronized prior to the acquisition. Hence, the most frequently sampled source (that is, process data) was down-sampled to match the less frequently sampled source (that is, UV data).

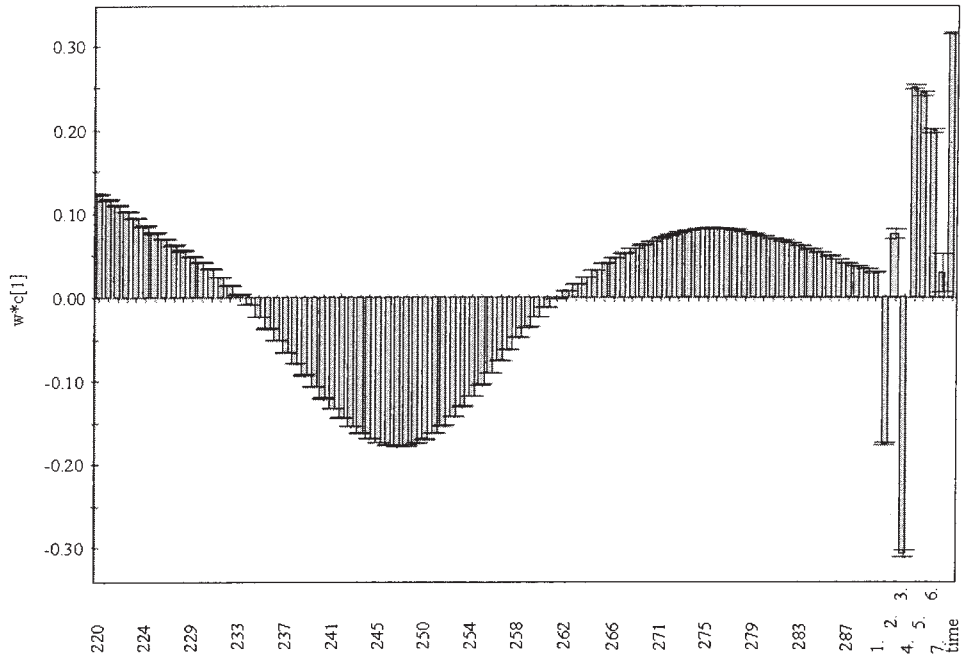

Figure 6. Plot for the w*c loadings of the first component of the combined model.

The first component uses 67% of the variation in **X** to explain 93% of the variation in the response with a Q^2 of 0.93. Every fifth wavelength is displayed. For process variable names, please refer to Figure 4. Time is used as the response variable in the model.

Time alignment in batch process studies is used to minimize the variability bias in batch duration and rate of progress. In this study the range of the model batches was 8 min (21% of average batch length). In addition, models with normalized time responses were also calculated. That did not affect the quality or the interpretation of the models and, hence, are not shown. The batches were aligned according to the implementation in SIMCA-P+ 11.0,[26] and time was used as response variable **y**.

Data matrices in batch processes are inherently three dimensional. In a process, K variables are measured at a specified interval. This results in a $J \times K$ matrix for each batch (J time points). For N batches there is an $N \times J \times K$ three-way data matrix, albeit it does not conform to multi-way analysis. Hence, the data matrix was unfolded according to Wold et al[4] to form an $(N \times J) \times K$ matrix.

Pre-processing and scaling

Due to a baseline shift, the spectroscopic profiles were Savitzky and Golay smoothed (5 points sub-model, quadratic polynomial) followed by taking the first derivative. This resulted in one component for the individual model for the spectral data.

The spectroscopic data were mean centered and process data were mean centered and scaled to unit variance. This scaling was used in the separate analyses of the respective data sets. Different block-scalings were tested to reach a feasible compromise between the differences in size and noise levels between the blocks of spectroscopic and process data. Using the base scaling of the individual models, the blocks were scaled to equal total sum of squares.

Results

Separate models for process and spectroscopic data

Separate batch models with time as the response variable were calculated for process and spectroscopic data respectively.

PLS Model for Process Data

The PLS model for process data has 3 components that use 75% of **X** to explain 86% of the variation in **y**, with a Q^2 value of 0.86. As expected, all of the design points deviate in the batch control charts due to deviating temperatures (Figure 4). PD11 appears normal both in scores and DmodX, and it is not until the temperature is manipulated with internal control after 15 min that the batch deviates. Thus, the low amount of catalyst is not detected. For PD12 the pattern is similar and the batch deviates after about 7 min when the stirrer is stopped. The fluctuations in reactor pressure cause PD12 to deviate considerably.

PLS Model for Spectroscopic Data

A one-component PLS model was calculated using the spectroscopic data using 93% of **X** to explain 89% of the variation in **y** with a Q^2 of 0.89. Again, all of the designed batches deviate in the batch control charts along with PD11, but this time it is due to deviating amounts of starting material and product (Figure 5). When DoE7 deviates after 4 min, it is because the amount of product is high. The trajectory of DoE9 is deviant for the duration of the batch, and when the batch finally goes outside the boundary it is because the amount of product is low. For PD11 variation that stems from temperature manipulations is discernable. The only process deviation that is

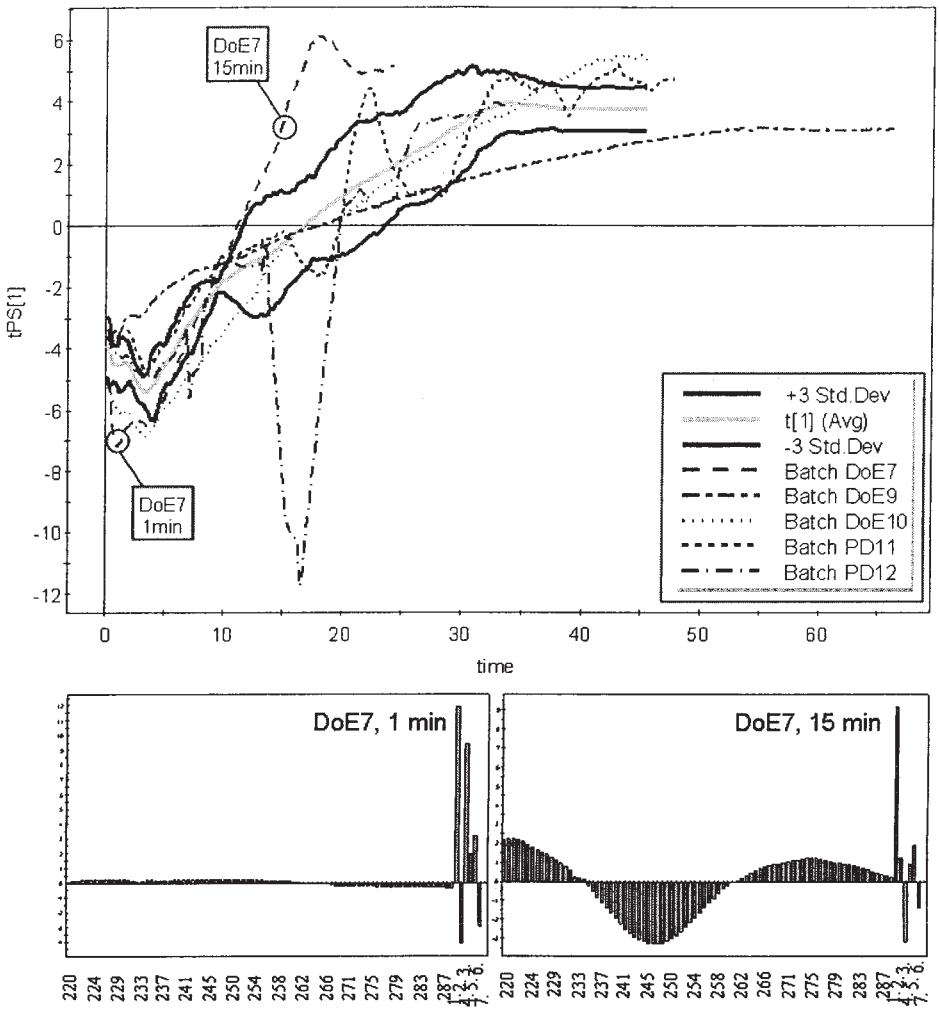

Figure 7. Predicted score values (above) for the batches that were not part of the model, together with the contribution to scores plot for batch DoE7 at 1 (lower left) and 15 min (lower right).

The boundaries correspond to average score value ± 3 standard deviations. For process variable names, please refer to Figure 4.

readily visible in the batch control chart is the effect of the stirrer stop for PD12, which otherwise appears normal for the duration of the batch. The decrease in reaction pressure is not detected.

Neither the model for spectroscopic data, shown in Figure 5, nor the model for process data, shown in Figure 4, has the ability to detect all deviations included in this study. Amounts of the starting material or product cannot be elucidated from process data alone. Spectroscopic data give only little information regarding the physical state of the process or possible malfunctions in equipment or sensors.

MSPC with both spectroscopic and process data

An MSPC was built using the combined data set with spectroscopic and process data and time as the response variable. The PLS model has three components that use 84% of the variation in **X** to describe 96% of the variation in **y**, with a cross validated Q^2 of 0.96.

The loadings of the batch model show that there is a strong contribution from many of the spectroscopic variables (Figure 6). Generally, after the process has run for a time there is a high content of product and low content of starting material. Process variables with a big influence on the progress of the batch are gas feed, jacket inlet and outlet temperatures, and flow rate of oil. The gas feed is low towards the end of the reaction as there is little or no starting material left. The temperatures are generally higher towards the end of the process compared to start-up conditions.

The batches of the design and the process deviation batches, the abnormal batches, all deviate in the batch control chart constructed from the first score vector t[1] (Figure 7). With DoE for initial concentration of starting material and reaction temperature together with carefully planned process deviations, all batches deviate for different reasons. DoE7 deviates immediately, and analysis of what contributes to the score values reveals a high reactor temperature and a high content of both starting material and product. When DoE7 is analyzed after 15 min, it is evident that the progression of the batch is rapid. The assessment of the batch would be that the high reaction temperature has increased the reaction rate to the point where the

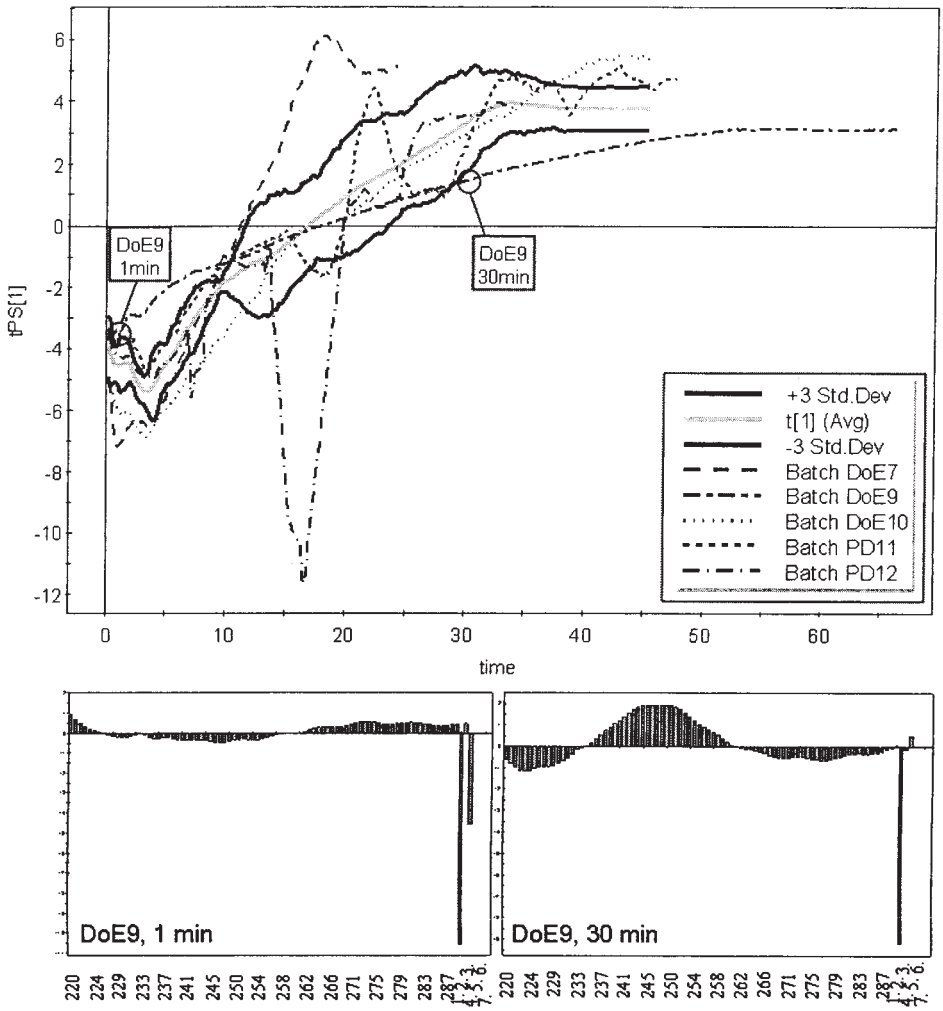

Figure 8. Predicted score values (above) for the batches that were not part of the model together with the contribution to scores plot for batch DoE9 at 1 (lower left) and 30 min (lower right).

The boundaries correspond to average score value ± 3 standard deviations. For variable names, please refer to Figure 4.

batch is way ahead of schedule. It is also possible that the initial concentration of the starting material is important for the reaction rate, which is seen in the contribution plot with high level of starting material from the onset of the reaction (Figure 7).

DoE9 deviates right from the start, and the contributing variables display a low content of starting material and a low reaction temperature (Figure 8). This results in the reaction progressing very slowly, which is easily seen in the batch control chart. The last of the DoE batches, DoE10, is very similar to DoE7 at the start of the process. This is due to a high content of starting material. The difference in reaction temperature between the batches becomes evident later in the process as the progression of DoE10 is slower. DoE10 is outside the boundaries in the first part of the process and after, for example, 10 min, this is because of a high content of starting material and product as well as low temperatures in the reactor and jackets (contribution plot not shown here).

Comparing Table 2 with Figure 7, it is obvious that the process manipulations that were performed on batches PD11 and PD12 cause visible changes in the batch control chart. PD11 is outside the boundary from the start and for most of the duration of the batch. The major contribution to the deviation in the first score vector comes from the UV data. This suggests that the fact that there is 10% less catalyst present in the reaction mixture might have an adverse effect on the chemical reaction as the spectra indicate a slower reaction. PD12 behaves like a normal batch with a couple of exceptions. When the stirrer is turned off after 7 min, the batch deviates and the reaction rate is slowed. After 13 min, the reaction pressure is decreased from 4 bars to 2 bars and the batch again goes outside the boundaries. When the reaction pressure is back up to 4 bars after approximately 20 min, the progression of the batch goes back to normal. Despite the fact that the stirrer speed manipulation is visible, the contribution plots (not shown here) display normal levels of starting material and product for the rest of the duration although it is very important to have information about process parameters deviation before it affects the chemistry itself.

Conclusions and Discussion

A wide variety of deviations were introduced in this batch process study. They were all detected by the combined data set

model, but not by using the UV spectroscopy or process data separately. A deviation due to, for example, low concentration of starting material and high reactor temperature can only be detected and interpreted using a combined data set approach. In more elaborate cases, with the possibility of interactions between chemical and physical factors, the results show that the sensitivity should be superior to that of individual models.

In accordance with the PAT initiative by the FDA, the focus of this study has been on using DoE and multivariate methods together with process analyzers in order to better understand the process. It has been demonstrated that a combination of spectroscopic and process data provides a deeper knowledge of the studied process. For the studied process, increased temperature speeds up the reaction. However, a high initial concentration of starting material leads to abnormal amounts of product, regardless of reaction temperature, implying that the kinetics rather than the thermodynamics of the reaction is of interest. Reducing the amount of catalyst by 10% significantly slows the reaction. The reaction is robust to fluctuations in stirrer speed, and stopping the agitator for 5 min does not significantly change the reaction rate or the outcome of the reaction. Temporarily reducing the pressure in the reactor does not seem to have an adverse effect on the reaction. The benefits of this approach should extend to other cases where there are more complex relationships between chemistry and physics.

The methods used are well known, and implementation to different batch processes in both primary and secondary production is feasible. A better understanding of the process together with process control tools should lead to a well-understood process and, ultimately, real time release.

Literature Cited

1. U.S. Department of Health and Human Services, Food and Drug Administration. *Guidance for Industry, PAT—A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*. 2004. http://www.fda.gov/cder/guidance/6419fnl.pdf. Accessed 01/12/2006.
2. Kourti T, MacGregor JF. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemom Intell Lab Syst.* 1995; 28:3-21.
3. Nomikos P, MacGregor JF. Multi-way partial least squares in monitoring batch processes. *Chemom Intell Lab Syst.* 1995;30:97-108.
4. Wold S, Kettaneh N, Fridén H, Holmberg A. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemom Intell Lab Syst.* 1998;44:331-340.
5. MacGregor JF, Kourti T. Statistical process control of multivariate processes. *Control Eng Practice.* 1995;3:403-414.
6. Andersson M, Folestad S, Gottfries J, Johansson MO, Josefson M, Wahlund KG. Quantitative analysis of film coating in a fluidized bed process by in line NIR spectrometry and multivariate batch calibration. *Anal Chem.* 2000;72:2099-2108.
7. Blanco M, Serrano D. On-line monitoring and quantification of a process reaction by near-infrared spectroscopy. Catalysed esterification of butan-1-ol by acetic acid. *Analyst.* 2000;125:2059-2064.
8. Box GEP, Hunter WG, Hunter JS. *Statistics for Experimenters—An Introduction to Design, Data Analysis, and Model Building* (1st ed.). New York: Wiley; 1978.
9. Eriksson L, Johansson E, Kettaneh N, Wikström C, Wold S. *Design of Experiments—Principles and Applications*. Umeå, Sweden: Umetrics AB; 2000.
10. Lundstedt T, Seifert E, Abramo L, Thelin B, Nyström Å, Pettersen J, Bergman R. Experimental design and optimization. *Chemom Intell Lab Syst.* 1998;42:3-40.
11. Jackson JE. *A User's Guide to Principal Components* (1st ed.). New York: John Wiley & Sons, Inc.; 1991.
12. Höskuldsson A. A combined theory for PCA and PLS. *J Chemometrics.* 1995;9:91-123.
13. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom Intell Lab Syst.* 1987;2:37-52.
14. Wold S, Geladi P, Esbensen K, Öhman J. Multi-way principal components- and PLS analysis. *J Chemometrics.* 1987;1:41-56.
15. Geladi P, Kowalski B. Partial least-squares regression: a tutorial. *Anal Chim Acta.* 1986;185:1-17.
16. Höskuldsson A. PLS regression methods. *J Chemometrics.* 1988;2: 211-228.
17. Martens H, Naes T. *Multivariate Calibration* (1st ed.). New York: John Wiley & Sons, Inc.; 1989.
18. Wold S, Trygg J, Berglund A, Antti H. Some recent developments in PLS modeling. *Chemom Intell Lab Syst.* 2001;58:131-150.
19. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemometrics.* 2002;16:119-128.
20. Barnes RJ, Dhanoa MS, Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl Spectrosc.* 1989;43:772-777.
21. Geladi P, MacDougall D, Martens H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl Spectrosc.* 1985;39:491-500.
22. Savitzky A, Golay MJE. Smoothing and differentiation by simplified least squares procedures. *Anal Chem.* 1964;36:1627-1632.
23. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. *Multi- and Megavariate Data Analysis—Principles and Applications*. Umeå, Sweden: Umetrics AB; 2001.
24. Kundu B, Sawant D, Partani P, Kesarwani AP. New application of Pictet-Spengler reaction leading to the synthesis of an unusual seven-membered heterocyclic ring system. *J Org Chem.* 2005;70:4889-4892.
25. *Matlab* [computer program]. Version 6.5.0. Natick: The MathWorks; 2002.
26. *Simca-P+* [computer program]. Version 11.0.0.0. Umeå: Umetrics AB; 2005.

Manuscript received Jan. 12, 2006, and revision received May 23, 2006.