JOURNAL OF CHEMOMETRICS
J. Chemometrics 2001; **15**: 665–683
DOI: 10.1002/cem.666

Statistical analysis of catalyst degradation in a semi-continuous chemical production process

Eleftherios Kaskavelis[1], Elaine Martin[1]*, Philip Jonathan[2] and Julian Morris[1]

[1]*Centre for Process Analytics and Control Technology, University of Newcastle, Newcastle-upon-Tyne NE1 7RU, UK*
[2]*Shell Research and Technology Centre—Thornton, PO Box 1, Chester CH1 3SH, UK*

SUMMARY

The effect of decaying catalyst efficacy in a commercial-scale, semi-continuous petrochemical process was investigated. The objective was to gain a better understanding of process behaviour and its effect on production rate. The process includes a three-stage reaction performed in fixed bed reactors. Each of the three reaction stages consists of a number of catalyst beds that are changed periodically to regenerate the catalyst. Product separation and reactant recycling are then performed in a series of distillation columns. In the absence of specific measurements of the catalyst properties, process operational data are used to assess catalyst decay. A number of statistical techniques were used to model production rate as a function of process operation, including information on short- and long-term catalyst decay. It was found that ridge regression, partial least squares and stepwise selection multiple linear regression yielded similar predictive models. No additional benefit was found from the application of non-linear partial least squares or Curds and Whey. Finally, through time series profiles of total daily production volume, corresponding to individual in-service cycles of the different reaction stages, short-term catalyst degradation was assessed. It was shown that by successively modelling the process as a sequence of batches corresponding to cycles of each reaction stage, considerable economic benefit could be realized by reducing the maximum cycle length in the third reaction stage. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: multivariate statistical modelling; semi-continuous process; catalyst decay

1. INTRODUCTION

Process analysis, monitoring and control rely on the availability of appropriate mathematical models to represent the system of interest. A common but sometimes very demanding approach is to develop a first-principles or mechanistic model of the process based upon knowledge of the chemical and physical phenomena underlying the process operation. Empirical data-based modelling is a widely used alternative to mechanistic modelling, since it requires less specific knowledge of the process being studied than that needed to develop a first-principles model. Empirical modelling techniques

* Correspondence to: E. Martin, Centre for Process Analytics and Control Technology, University of Newcastle, Newcastle-upon-Tyne NE1 7RU, UK.
E-mail: e.b.martin@ncl.ac.uk

Contract/grant sponsor: Statistics Group, Shell Research Ltd
Contract/grant sponsor: Foresight Centre for Process Analytics and Control Technology, University of Newcastle, UK

Received 30 September 2000
Accepted 16 November 2000

require data (measurements) collected on those variables believed to be representative of process behaviour and of the quality or properties of the product or system output. Statistical regression techniques are now routinely used in the process industries for building empirical models.

In practice, a model is an approximate representation of a real world system, and model building is a balance between model simplicity, model accuracy and computational time. Numerous regression techniques are available that differ in terms of complexity and speed of computation. These include linear regression, linear regression through data transformations, non-linear regression, non-parametric regression analysis and neural networks. Each has its own advantages and disadvantages. As a result, practitioners are required to select the most appropriate tool according to the modelling objective and the required simplicity and accuracy of the application.

The aim of the paper is to demonstrate that through the structured application of chemometric modelling tools to a complex industrial process, process features of significant manufacturing importance can be identified from non-standard data. Data from a major petrochemical process for the production of higher alkenes (olefins) form the basis of the study. More specifically, the effects of catalyst decay in the three-stage reaction section are investigated. The three-stage reaction comprises a number of fixed bed reactors. Frequent catalyst replenishment by regeneration or replacement is required. However, through the interchanging of so-called ‘spent’ beds with essentially identical but replenished beds, production is maintained. The process can therefore be viewed as being of a semi-continuous form. Following the reaction section, products are separated from reactants (which are recycled) and extracted in a series of distillation columns.

Within a single cycle of bed operation, catalyst efficacy decays over time. However, the decay profile cannot be quantified easily owing to multiple competing sources of process variation. In the absence of data to characterize catalyst status directly as a function of time, statistical methods are used to identify the relationship between catalyst status and process performance. Catalyst in the first and third reaction stages is regenerated, whilst the catalyst in second-stage beds is replaced. Since regeneration is never 100% effective, a long-term efficacy decay effect is present. Isolation of this trend is difficult owing to confounding with other sources of variation, most notably time. A simple averaging technique, based on the analysis of time series profiles of multiple cycles, is used to quantify short-term (within-cycle) catalyst effect and demonstrate that the performance of the process, in economic terms, can be improved by reducing the operational cycle lengths in one of the three reaction stages.

A number of statistical tools were applied to develop predictive models of production flow as a function of process operation and short- and long-term catalyst decay. Multiple linear regression (MLR) has been extensively used in the development of predictive empirical models. However, when dealing with highly correlated multivariate problems, the traditional approach of MLR can lead to singular solutions or very imprecise parameter estimation [1]. These issues can be overcome by applying alternative regression methodologies such as the regularization techniques of ridge regression [2] and partial least squares (PLS) [3]. Ridge regression overcomes the ill-conditioning problem, whilst PLS not only addresses the collinearity problem but also reduces the dimensionality of the problem and can provide a filtering tool for measurement noise.

Although the overall objective of the study was not to carry out an extensive comparison of different methods, the above approaches were compared with Curds and Whey [4] and the non-linear technique of non-linear partial least squares to assess the applicability of the different tools when applied to industrial data. Curds and Whey is a method for predicting several response variables from the same set of explanatory variables. The advantage of this approach is that it takes account of the correlations between the response variables to improve prediction accuracy. In practice, when dealing with real and complex chemical and physical systems, linear techniques cannot reliably be used to model the underlying structure, as it may exhibit significant non-linear characteristics. Of particular

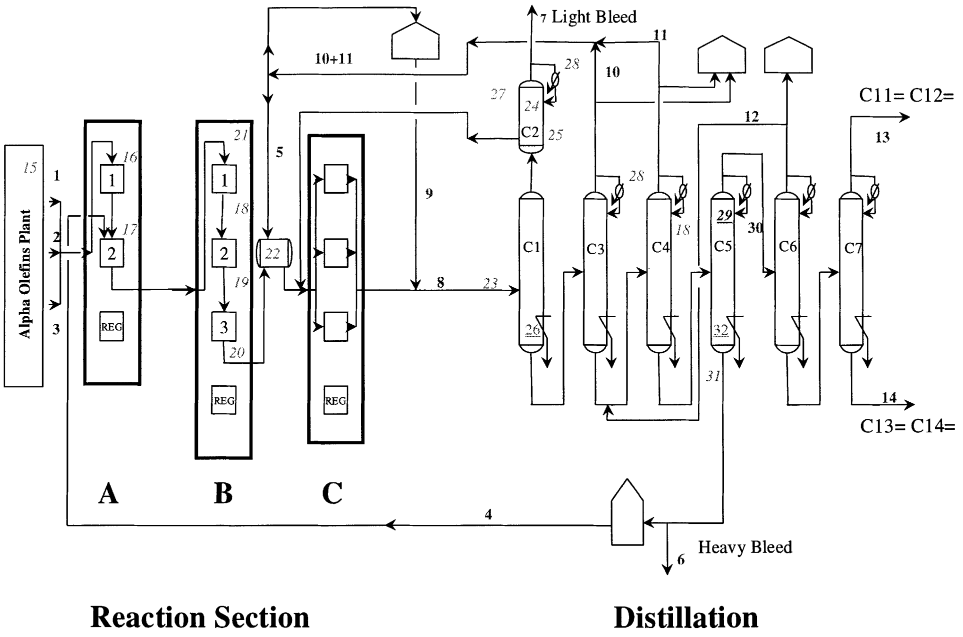

Figure 1. Schematic diagram of the process.

interest in this paper is to model the non-linearities inherent within the data using non-linear PLS. A number of algorithms have been proposed to integrate non-linear features within the linear PLS framework to produce a non-linear PLS algorithm that retains the statistical properties of the linear methodology [5–9]. Wold *et al.* [5] proposed a polynomial (quadratic) PLS algorithm that retains the framework of linear PLS but modifies the relationship between the predictor and the response latent variables to be non-linear. More recently a novel error-based PLS approach has been developed by Baffi *et al.* [6], and this algorithm has been used here.

2. THE PROCESS

The process (Figure 1) consists of two sections, reaction and distillation. The reaction section comprises a series of three catalytic reaction stages, A, B and C, which are operated in predetermined combinations of the fixed bed reactors. The second stage, the distillation section, consists of a network of seven distillation columns in which separation of reaction products occurs. A number of recycle flows return approximately 80% of the total mass from the distillation section back to the reactors. Bleed streams remove unwanted side-products and contaminants. Feed enters the reaction section at approximately 15 t h^{-1}. The distillation section separates products, bleeds (both of which exit the process) and unprocessed reactants (which are recycled) at approximately 10, 2 and 130 t h^{-1} respectively.

The process converts unsaturated hydrocarbons in the form of primary alkenes of short or long carbon chain length to higher (secondary, tertiary, etc.) alkenes of intermediate chain length which are of higher economic value. Reaction stage A is responsible for purification of the reaction mixture. In stage B, specific isomerization reactions take place to create secondary, tertiary and higher alkenes from the primary alkene feed. In stage C, exchange reactions of the form $X=Y+Z=W \rightarrow X=Z+Y=W$ take place. Here standard chemical notation is adopted, with '=' referring to a

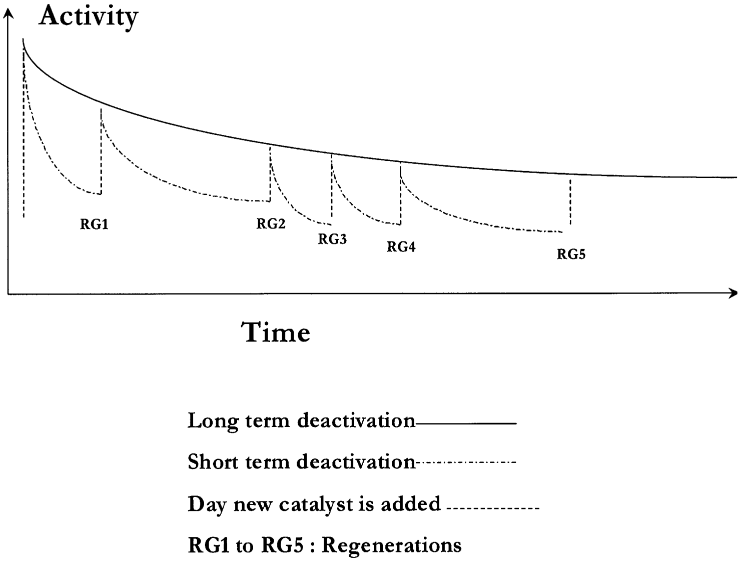

Figure 2. Catalyst efficacy for reactor stages A and C.

double bond and W, X, Y and Z representing hydrocarbon chains. In stage C the short-chain alkenes $(X = Y)$ react with long-chain alkenes $(Z = W)$, creating the desired products of intermediate chain length.

In the reaction section a total of eight reactor beds are operational at any one time. Catalyst efficacy in each bed decays over a period of days or weeks within each reactor, necessitating frequent bed regeneration or catalyst replacement, depending on the type of reactor. To maintain production, spent beds are exchanged for duplicate beds with regenerated catalyst for the A and C beds or new catalyst for the B beds. A bed is typically on-line for a cycle of between 1 and 3 weeks. In stage A, two beds are used on-line in series, whilst the catalyst in the third bed is regenerated. The reaction mixture enters the first bed (1 in Figure 1), moves onto the second bed (2 in Figure 1) and then exits to reaction stage B. The catalyst in the first bed is older than that in the second bed. When the catalyst in the first bed is considered to be spent, a cyclic rotation of the three beds is performed. The first bed is regenerated, the second bed becomes the first bed and the off-line regenerated bed becomes the second bed, thereby regulating overall catalyst efficacy in reaction stage A and maintaining production. In stage B a similar operating procedure is used, except that three beds are used on-line in series at any one time. Concurrently, fresh (new) catalyst is supplied to a fourth, off-line, bed. The beds in stage C are operated in parallel, usually with three beds on-line whilst a fourth is being regenerated.

Figure 2 illustrates the likely behaviour of catalyst efficacy over time for reaction stages A and C. Within a cycle, catalyst activity decays over a number of days until a decision is made to exchange a spent bed with a regenerated duplicate, thereby initiating a new cycle. Since regeneration is never totally complete, catalyst performance at the start of each cycle also decays slowly with time. For reaction stage B, fresh catalyst is used to replenish spent beds, hence no long-term trend is anticipated in this particular unit operation. Since fresh catalyst is expensive, there is strong economic incentive to run each B bed cycle for as long as possible before replacing catalyst. In contrast, catalyst regeneration (in stages A and C) is relatively inexpensive. Depending on the effect of catalyst decay on process throughput, economic consideration might support the lengthening of cycles for stages A and C.

Table I. Descriptions of the process variables used: F (flow), T (temperature), D (duty), P (pressure) and time variables present

Number	Variable description	
1	F: Light feed	Process
2	F: Medium feed	Process
3	F: Heavy feed	Process
4	F: Major recycle	Process
5	F: Recycle C1–4 to C	Process
6	F: Heavy bleed	Process
7	F: Light bleed	Process
8	F: Feed from ABC to distillation section	Process
9	F: Feed from tank to distillation section	Process
10	R: Recycle from C3 to C	Process
11	F: Recycle from C4 to C	Process
12	F: Recycle from C6 to C4	Process
13	F: C11= & C12=product	Product
14	F: C13= & C14=product	Product
15	T: Feed temperature to reactors	Process
16	T: A bed temperature position 1	Process
17	T: A bed temperature position 2	Process
18	T: B bed temperature position 1	Process
19	T: B bed temperature position 2	Process
20	T: B bed temperature position 3	Process
21	T: Drop pre first B bed	Process
22	T: Vessel between B and C	Process
23	T: Feed to distillation section	Process
24	T: Top of C1	Process
25	T: Bottom of C1	Process
26	D: Reboiler duty of C1	Process
27	T: Top vapour temperature of C2	Process
28	T: Cooling water C2 + C3	Process
29	P: Top of C5	Process
30	F: Reflux at top of C5	Process
31	T: Bottom of C5	Process
32	D: Reboiler duty C5	Process
33	Time: Days C bed configuration live	Process
34	Time: Days B bed configuration live	Process
35	Long-term catalyst decay	Process
36	F: Total product flow (equal to the sum of variables 13 and 14)	Product

Total production volume is determined by the efficacy of catalyst in each of the three reaction sections. As catalyst performance decays, product flows reduce and average retention time in the process increases. To compensate for this effect, the separating ability of the distillation columns can be improved by increasing their duties (energy consumption). The duties on columns 1, 2 and 5 in particular are varied by engineers on a routine basis. The operation of the remaining columns is not varied significantly. Consequently, overall feed flow rates through the plant, which are limited by the capacity of the reaction section, are also reduced.

3. THE DATA

The processing operation to be analysed is extremely complex and the data collected are non-homogeneous and therefore present a major challenge to chemometric techniques. Daily average

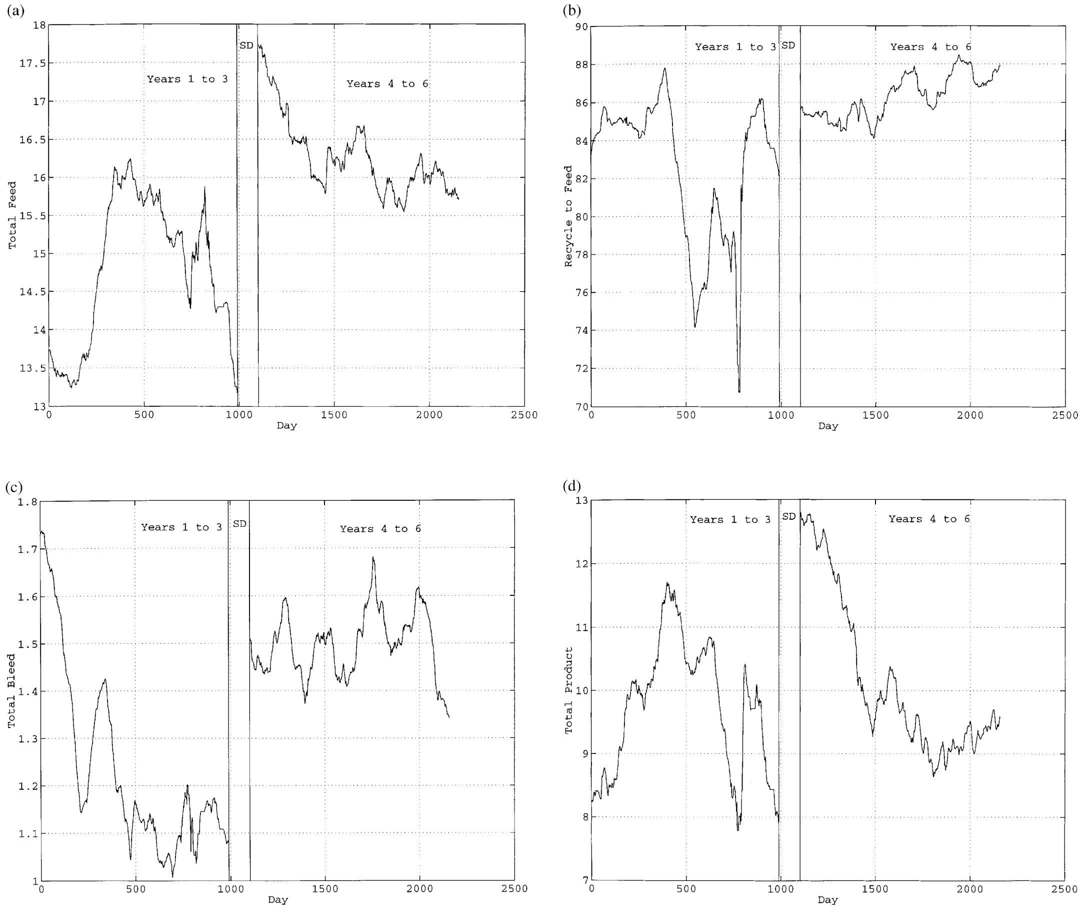

Figure 3(a). Filtered data: total feed, (b) recycle to feed, (c) total bleed, (d) total product.

operating data over a period of 6 years (April 1991–February 1997) were considered. The data set consisted of 2156 observations on 33 process variables characterizing feed, bleed, recycle and product streams as well as many other intermediate flow rates, reactor bed temperatures, pressures and distillation column conditions, and three product quality variables relating to product flow. Although no direct measurements on catalyst properties were available, three surrogate variables were defined. Variables 33 (days C bed configuration live) and 34 (days B bed configuration live) relate to short-term catalyst decay, whilst variable 35 acts as a surrogate for long-term decay. The complete set of variables is listed in Table I and shown on the process diagram (Figure 1). Following consultation with the process engineers, samples corresponding to periods of process shutdown and non-typical operation were removed from the data set, resulting in a revised data set comprising 1764 samples. These formed the basis of the subsequent exploratory analysis.

4. EXPLORATORY ANALYSIS

The first stage in the analysis was to pre-process the data. This included the identification, interrogation and handling of missing data and spurious data points, and the elimination of noise from

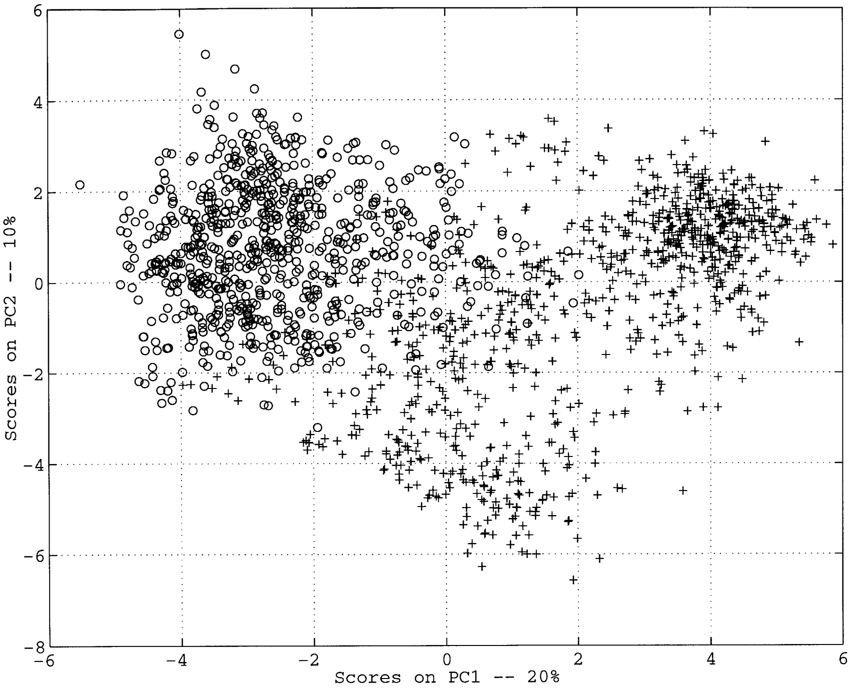

Figure 4. Score plot of principal components 1 and 2 (○, years 1–3; +, years 4–6).

process variables with a low signal-to-noise ratio. With respect to missing observations, one of three approaches can be adopted, sample or variable elimination, data imputation, or the use of algorithms that allow model building in the presence of missing values [10]. The former approach of elimination was selected. However, care needs to be taken in the elimination of measurements from data. If time delays are present in the process owing to recycles, residence times, etc., the data need to be time aligned to ensure that relationships between variables are not masked by the time shift effect. This was not an issue here, since the data were presented in terms of daily averages, and time delays were known to be of the order of hours and not days. Thus the issue of time alignment did not need to be addressed. However, variables with in excess of 30% of the original observations missing were removed from the analysis. This resulted in the removal of heavy bleed (variable 6) from the modelling data set, giving a reduced data set comprising 1540 complete observations on 32 explanatory (process) variables and three product quality variables.

Spurious points were not identified in the data owing to the presence of a data reconciliation algorithm within the distributed control system (DCS). The variables total feed (the sum of variables 1–3), total bleed (the sum of variables 6 and 7), total recycle (variable 4) and total product flow (variable 36) had low signal-to-noise ratios, thus it was decided to apply a moving average filter. The filtered profiles are given in Figures 3(a)–3(d). The size of the window (± 3 months) was selected to ensure that the key features of process behaviour were not lost owing to over-smoothing and that major trends were not significantly masked by process noise. From the figures it is clear that process operation in years 1–3 differs from that in years 4–6. In years 1–3 a number of significant process and operational changes were made at various times. For years 4–6, however, following a major shutdown (SD in Figures 3(a)–3(d)), process operation is more consistent; total feed (Figure 3(a)) and total product (Figure 3(d)) decrease smoothly until an approximately constant value of the mean is reached, whilst total bleed is approximately constant (Figure 3(c)).

The final stage before applying principal component analysis (PCA) and the modelling tools described in Section 5 was to standardize the data to zero mean and unit variance. PCA [11] was applied to the process data as defined in Table I. The first two principal components explained 33% of the variance of the sample correlation matrix. Interrogating the loading plot for principal component 1 (not shown), there was a clear indication that the temperature variables in stages B and C of the reaction section have the largest absolute contribution, whilst the variables dominating principal component 2 were the flows in the distillation section. From the score plot of principal components 1 and 2, Figure 4, the difference between years 1–3 (○) and years 4–6 (+) was confirmed. Based on this information, it was decided to only use the data for years 4–6 for the modelling work reported in Section 5, since from an operational perspective this is believed to correspond to a period of consistent operation. In contrast, data for the whole of the 6 year period were used for the estimation of the within-cycle production decay profiles discussed in Section 6. This is again justifiable from a process engineering perspective, since the short-term, within-cycle, trend is expected to be present for all operating conditions.

Finally, the existence of linear relationships between the process and quality variables was identified from the Spearman rank correlation matrix. From this analysis a number of high correlations (in excess of 0·7) were identified. These included the operating temperatures in the A and B beds (variables 17–20), the recycles from columns C3 and C4 and the individual product flows from C7 (variables 11–14) with the operating parameters for column C1 (variables 24–26), C bed cycle time (variable 33), long-term catalyst decay (variable 35) and total production (variable 36). These are important observations for the modelling work described in the next section.

5. MODELLING OF THE RATE OF PRODUCTION

In deciding on the modelling approaches to explore, a number of issues were considered.

- The process data are highly correlated (Section 4). Thus modelling techniques which handle multicollinearity (i.e. non-orthogonality of predictors), such as partial least squares or ridge regression, should be considered.
- The relationship between product flow and process operation may be non-linear. Thus a non-linear methodology should be investigated.
- Engineers on the plant would ideally like a simple model to implement within their process manufacturing software. Parsimony suggests that the simplest model consistent with the data should be identified. Multiple linear regression (established via a stepwise search procedure) is attractive for this purpose.
- The product flow can be partitioned into two streams with correlated flow rates. There is therefore the potential to model the individual product streams (although this is not a primary objective). The technique of Curds and Whey [4], which is specifically designed for this situation, was therefore considered.

For these reasons, the modelling techniques of partial least squares (linear and non-linear), ridge regression and stepwise multiple linear regression were considered. These methodologies were supplemented by Curds and Whey, an approach specifically developed for multiple responses.

Cross-validation [12] was used throughout the study to estimate the predictive performance of the methods requiring tuning (namely ridge regression, partial least squares and Curds and Whey). A common cross-validation strategy was adopted for all analyses. The data were partitioned into 30 groups of size 18. Each cross-validation group represents a contiguous interval in time of 18 days duration. This is an important consideration in modelling multivariate, autocorrelated time series. In particular, selecting cross-validation groups at random (with respect to the time order of the data)

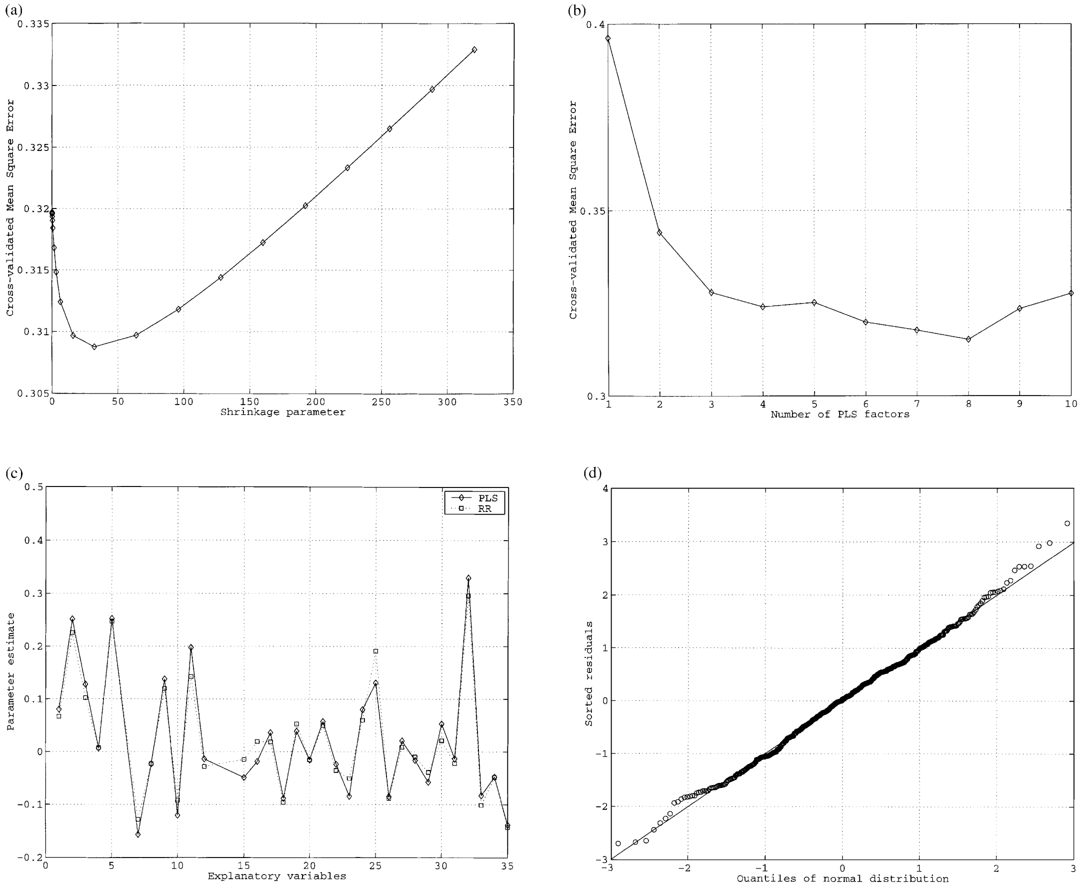

Figure 5(a). Variation in mean square error with shrinkage parameter for ridge regression. (b). Variation in mean square error with latent variables for partial least squares. (c). Parameter vectors for ridge regression (RR) and partial least squares (PLS). (d). Quantile–quantile plot for ridge regression.

produces over-optimistic estimates of future predictive performance. Since cross-validation is used to tune the models, the corresponding estimates for predictive performance will themselves be biased. A two-deep cross-validation strategy [13] would be necessary to provide unbiased estimates in these situations. Predictive performance is reported in terms of cross-validated mean square error (cvMSE). Since the response is standardized, this is related to the cross-validated coefficient of determination (cvR^2) by cvMSE + cvR^2 = 1.

5.1. Ridge regression and partial least squares

Ridge regression is a well-established technique for multivariate modelling introduced by Hoerl and Kennard [2]. By adding a constant (the 'shrinkage' or 'ridge' parameter) to the diagonal of the sample covariance matrix in multiple regression, stable but biased regression results. In ridge regression, cross-validation was used to estimate the shrinkage parameter. In this application a search was carried out over a pre-specified grid [13]. A plot of cross-validated mean square error against ridge parameter is shown in Figure 5(a). The value of ridge parameter that gives the best predictive performance is 30 (corresponding to approximately the trace of the sample covariance matrix), giving a cross-validated mean square error of approximately 0·31 (corresponding to a cross-validated R^2 of 0·69).

Partial least squares is a well-reported technique for linear modelling in the presence of multicollinearity [3]. In partial least squares, cross-validation is used to select the optimal number of latent variables to retain. A plot of cross-validated mean square error is given in Figure 5(b). The figure shows that four factors correspond to a cvMSE of 0·32. These factors explain a total of approximately 43% of the total variance of the explanatory data and 75% of the response variance. For this application it can be seen that the predictive performance of partial least squares is almost identical to that of ridge regression. Moreover, the regression parameters estimated by partial least squares and ridge regression are very similar, as illustrated in Figure 5(c). The explanatory variables having the greatest individual contributions to the regression vector are the medium feed rate (2), the light bleed (7) and the two recycles (5 and 11), the bottom temperature for column 1 (25), the duty of distillation column 5 (32) as well as the long-term catalyst decay (35). Although the effect of intercorrelation of explanatory variables needs to be borne in mind, the key variables identified can be related to process knowledge, and thus it was concluded that the model was physically realizable. A quantile–quantile (QQ) plot of the sorted residuals for ridge regression against expected normal order statistics is given in Figure 5(d), indicating that the assumption of normality is satisfied. A similar plot was obtained for PLS (not shown).

5.2. *Stepwise multiple linear regression and response surface fitting*

A stepwise linear search procedure was used to identify multiple linear models using the S-PLUS software [14]. A number of competing models of varying complexity were identified with multiple R^2 of approximately 0·7. Many of these models included those variables found to be influential in the ridge and partial least squares models. The final set comprised variables 2, 5, 7, 11, 25, 32, 33 and 35. The only additional variable identified compared to ridge regression and PLS was variable 33 (days C bed configuration live) which relates to short-term decay. Using this subset of variables, a full response surface model was fitted (including pairwise interactions and quadratic terms). Results indicated that the interactions of variable 25 (bottom temperature for column 1) with variables 2 (feed rate), 5 (recycle) and 7 (light bleed flow) might be worth adding to the original linear model. However, the improvement in the predictive ability of the model was minimal compared to the added complexity of the model which would make it less acceptable for the engineers to utilize and understand. Thus these interactions were not included in the final model. Parameter estimates for the nine-term multiple linear regression model are given in Table II. Comparing these values with the parameter estimates for ridge regression and PLS, there is a strong similarity between the two sets. Thus it appears that ridge regression, PLS and MLR provide comparable models, as previously discussed by Frank and Friedman [15].

5.3. *Non-linear partial least squares*

Wold *et al.*, [5] proposed a non-linear PLS algorithm which retained the framework of linear PLS but used second-order polynomial (quadratic) regression (QPLS) to fit the functional relation between each pair of latent variables. In their paper the authors identified the main drawbacks of merging non-linear regression techniques within the framework of the linear PLS algorithm. They proposed a solution and went on to show how their approach to updating the input weights is suitable for any continuous and differentiable functional relationship between the input and the output scores. Based on this work, Baffi *et al.* [6] proposed an error-based updating procedure which resulted in a new quadratic PLS algorithm. This was shown to provide improved modelling capabilities over the algorithm of Wold *et al.* [5]. The approach of Baffi *et al.* [6] was adopted in this work. Using the

Table II. Parameter estimates for the stepwise selection multiple linear regression model

Variable	Parameter estimate	Standard error	*t* statistic
2	0·1730	0·0252	6·8687
5	0·1602	0·0352	4·5453
7	−0·1834	0·0253	−7·2472
11	0·2149	0·0510	4·2102
25	0·1548	0·0532	2·9116
32	0·3217	0·0507	6·3450
33	−0·1238	0·0243	−5·0956
35	−0·1017	0·0415	−2·4510

cross-validation strategy described, it was found that only one non-linear partial least squares component should be retained, explaining approximately 16% of the explanatory variance and about 80% of the response variance. The corresponding cross-validated mean square error of 0·32 suggests that no additional benefit was realized through the use of non-linear PLS. Including additional non-linear latent variables resulted in an increase in the cvMSE with only a minor increase in the percentage variability explained (Figure 6). For example, for two latent variables the *x*-variability explained was 18%, whilst for the quality variables 83% of the total variability was explained. It was thus concluded that there was no real improvement in the predictive ability of the model by incorporating non-linear terms into the analysis.

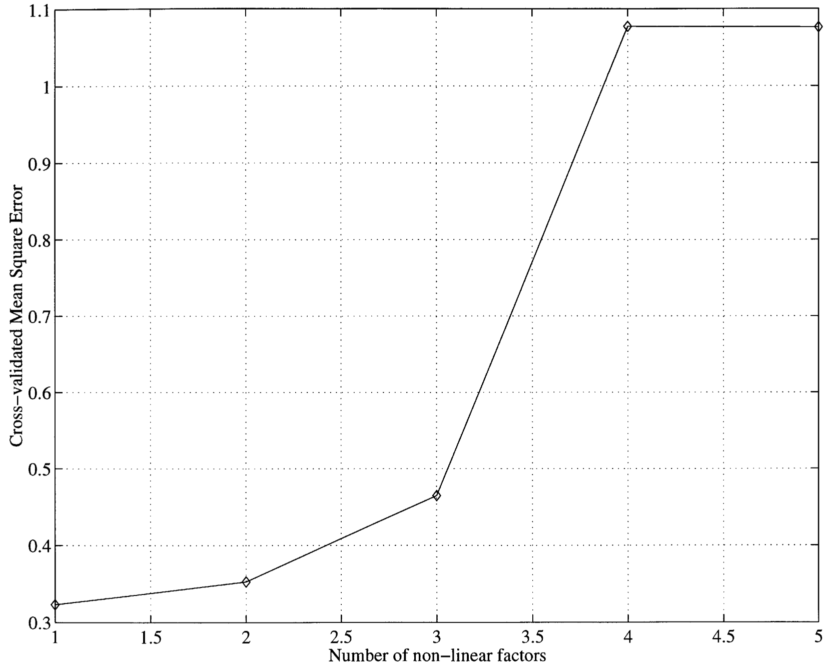

Figure 6. Variation in mean square error with non-linear partial least squares factors.

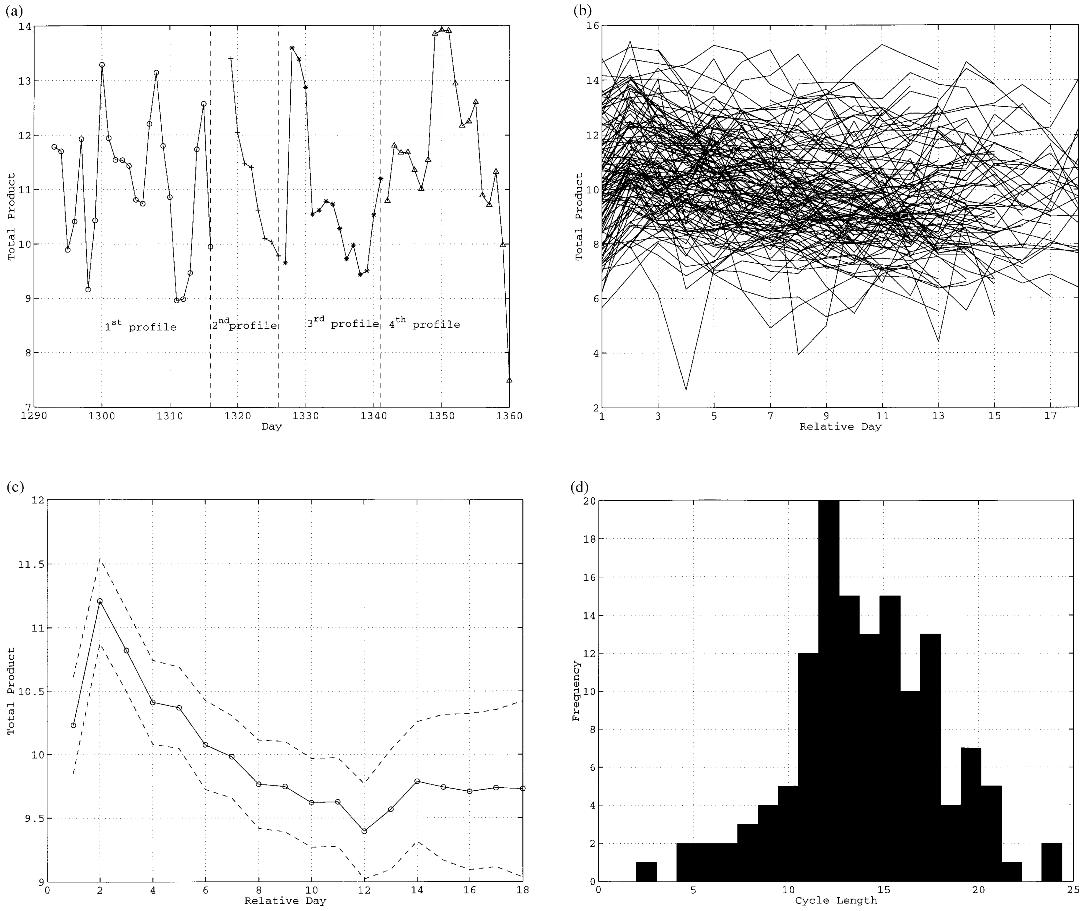

Figure 7(a). Time series plots of product flow for individual cycles in reaction stage C. (b). Time series profiles of product flow for C cycle. (c). Averaged time series profiles with confidence intervals for C cycle. (d). Histogram of cycle lengths.

5.4. Curds and Whey

Curds and Whey [4] is a method for the biased prediction of a multivariate response, designed to take specific advantage of the correlation between response variables. The method can be viewed as a form of canonical correlation analysis applied to the individual ridge predictands (with a common ridge parameter) to improve the overall predictive performance. Again it was assessed using cross-validation. Generalized cross-validation can be used to estimate the shrinkage parameter in the canonical analysis part, thus there is only a need for cross-validation to select the common ridge parameter.

The method was applied to the two product streams C11=/C12= and C13=/C14= (variables 13 and 14 respectively) of which the total product flow is composed. The sample intercorrelation of these variables was 0·26. Previous analysis of the individual product streams using partial least squares and ridge regression had suggested that the C11=/C12= stream was more amenable to modelling. This was confirmed by the Curds and Whey analysis, which yielded good predictive performance for

C11=/C12= (cvMSEs of approximately 0·3) but much poorer performance for C13=/C14= (cvMSEs of approximately 0·7). The Curds and Whey predictors were observed to be very similar to those based on the individual responses. It was therefore concluded that in this case Curds and Whey analysis does not offer any additional benefit with respect to modelling the total product flow.

6. ESTIMATION OF WITHIN-CYCLE PRODUCTION DECAY AND OPTIMIZATION OF CYCLE LENGTH

The effect of within-cycle catalyst degradation on process performance was estimated by assuming the process to be a contiguous sequence of batch runs where a batch corresponded to a cycle of operation in any of the reaction stages A, B or C. By averaging over all cycles for a given reaction stage, an estimate of the expected decay of product flow within a given cycle is obtained. It was assumed that averaging reduced the effect of other sources of variability in relative terms, and thus the decay profile could be attributed to short-term catalyst degradation. From a process engineering perspective this assumption was valid.

The analysis is first outlined before describing the procedure in more detail. Figure 7(a) shows typical time series plots of product flow for four individual cycles in reaction stage C. The individual cycles can be isolated from the process operational information relating to bed switching times. Thus profiles of product flow as a function of time since the start of the present C cycle can be identified (Figure 7(b)). The next step was to average these profiles at each time point and calculate the point wise 95% confidence band for the mean trajectory (Figure 7(c)).

As mentioned previously, product flow decreases with increasing catalyst age; thus, if only one of the A, B or C beds were switched, production profiles as seen in Figure 2 would be realized. However, some of the profiles in Figure 7(b) (e.g. for C) contradict this trend (i.e. appear to increase over time). This is because switches in other beds (A and B) are occurring during the C cycles, thus causing product flow to increase. Averaging over the cycles is therefore analogous to estimating the main effects in a linear model.

The mean profile for the C cycles, Figure 7(c), indicates that production rate drops from approximately 11·2 to approximately 9·7t hr^{-1} over the first 10 days of the cycle, corresponding to a daily reduction of some 0·15t $\mathrm{hr}^{-1}\mathrm{day}^{-1}$. It is interesting to compare these rates with those found from the modelling in Section 5. The parameter estimate for variable 33 (days C bed configuration live) from Figure 5(c) is $-0{\cdot}10$ with respect to the standardized variables. Correcting for the standard deviation of total production ($\sim$5·2) and C cycle day ($\sim$3·4), the daily reduction from the model is also approximately $-0{\cdot}15$ ($= -0{\cdot}1 \times 5{\cdot}2/3{\cdot}4$). Thus, based on the parameter estimate for variable 33 (days C bed configuration live) from the linear model in Table II, a slight decay is indicated, confirming what is seen in Figure 7(c).

Since cycle lengths are not constant, as seen in Figure 7(d), the confidence band for the mean profile increases with increasing cycle time, i.e. there are fewer longer cycle lengths. Furthermore, since daily average data are being considered, there is some ambiguity concerning the first day of a cycle. The convention adopted is that the first day of a cycle is defined as that day when the new or regenerated catalyst was first used. However, in practice the first full day of that cycle is actually the following day (day 2 in Figure 7(c) for this example). This explains the odd profile for relative days 1 and 2 in the figure.

Mathematically, let $\{p(i)\}$, $i = 1, \ldots, n$, be the time series of daily average product flow, where n is the total number of available days, i.e. 1764 samples. Let $\{T_i^{\mathrm{C}}\}, i = 1, \ldots, N^{\mathrm{C}}$, be a vector denoting the first day of each cycle, shown by the vertical lines in Figure 7(a), and let N^{C} be the number of cycles within time period n in reaction stage C (i.e. the length of the vector). Cycle profiles $\{x_{ij}^{\mathrm{C}}\}, i = 1, \ldots, N^{\mathrm{C}}, j = 1, \ldots, L_i^{\mathrm{C}}$, shown overlaid in Figure 7(b), are constructed according to the

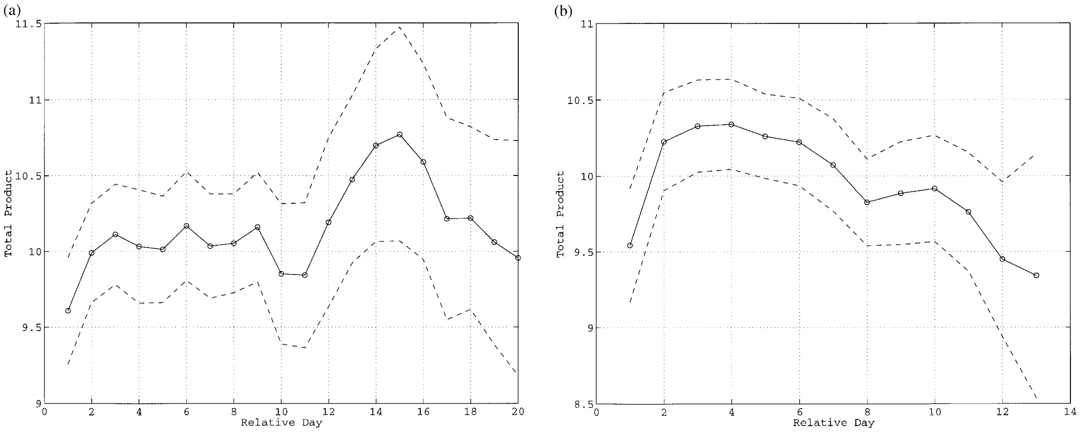

Figure 8(a). Averaged time series profiles with confidence intervals for A cycle. 8(b). Averaged time series profiles with confidence intervals for B cycle.

relationship

$$\{x_{ij}^{\mathrm{C}}\} = p(T_i^{\mathrm{C}} - 1 + j), \quad j = 1, 2, \ldots, L_i^{\mathrm{C}}, \quad i = 1, 2, \ldots, N^{\mathrm{C}} \tag{1}$$

where L_i^{C} is the length of cycle i, given by $T_{i+1}^{\mathrm{C}} - T_i^{\mathrm{C}}$; that is, the difference between the first day of the next cycle and the first day of its previous one. Since this difference is variable, the cycle profiles can be thought of as being batch processes of unequal length. The mean cycle profile Figure 7(c) with respect to the reaction cycles of stage C is therefore $\{\bar{x}_{\cdot j}^{\mathrm{C}}\}, j = 1, 2, \ldots, L^{\mathrm{C}}$.

It is clear from Figure 7(b) that there is considerable variability between individual cycle profiles. Notably, since cycle lengths for reaction stages A, B and C vary considerably (with cycle lengths of 5–30 days for A and 4–18 days for B), the relative times of A and B cycle switches, with respect to C cycle switches, also change. This is an additional complication in the data analysis, since bed switches that occur close in time (i.e. on the same day) might have a greater impact on the overall process performance than switches that occur further apart in time. Furthermore, since catalyst degradation effects are anticipated by the process chemists, various operational steps are taken, particularly in the distillation section, to maintain production levels.

Similar analyses were performed for reaction sections A and B. The results are illustrated in Figures 8(a) and 8(b) respectively. The effect of short-term catalyst degradation is clear for stage B, where product flow can be seen to decrease on average by approximately 0·7 t h^{-1} over a cycle. However, for stage A, catalyst degradation effects are not apparent. Discussions with process chemists confirmed that the effects of catalyst degradation are known from experience to be more pronounced in reaction section C than in sections B and A. The results of the data interrogation agreed with their process understanding.

Figures 7(c) and 8(b) suggest that reducing cycle times for reaction stages B and C would result in increased total product throughput. However, to operate with shorter cycles, more frequent catalyst regenerations (bed C) or fresh catalyst supplies (bed B) would be required, thereby increasing operational costs. The viability of reducing cycle lengths must therefore be assessed in terms of the competing economic factors. In reality, one bed regeneration for reaction stage C is relatively inexpensive, being of the order of £5000, implying that reducing C cycle lengths is worthy of consideration. In contrast, the fresh catalyst used for reaction stage B is expensive. A bed switch in

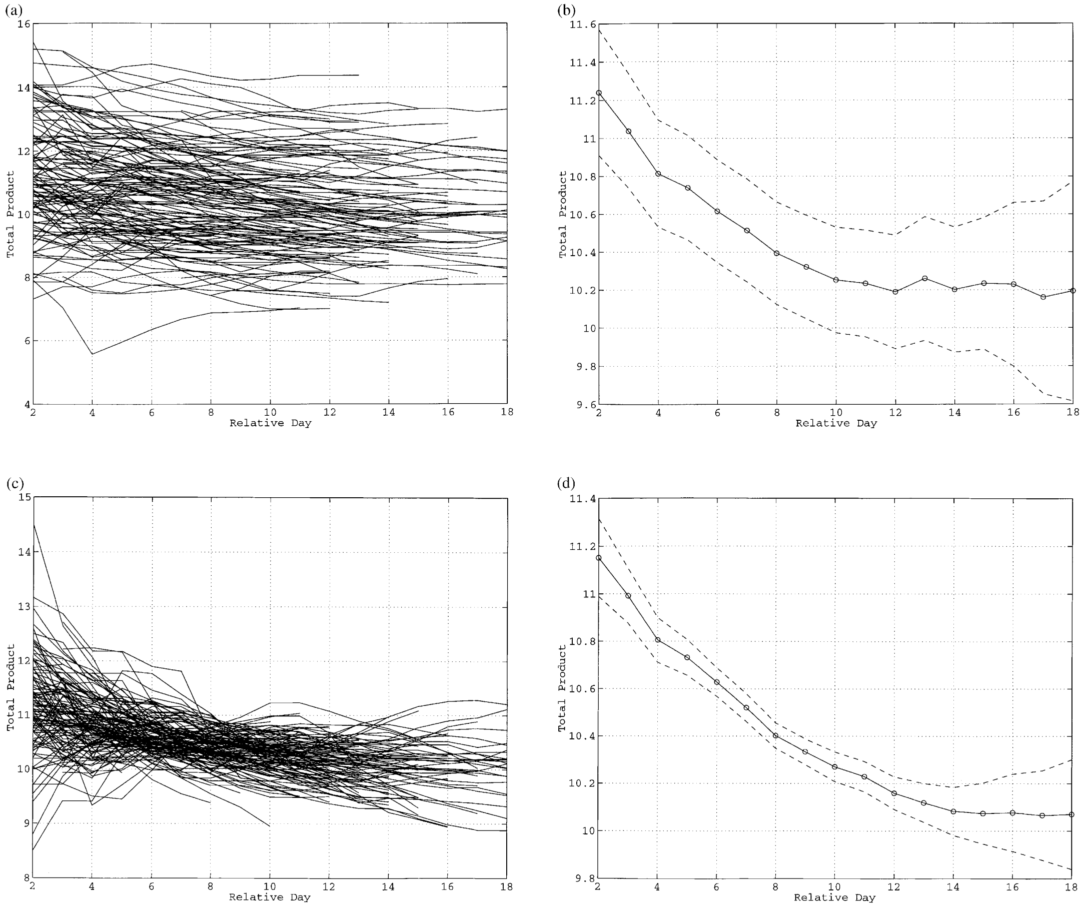

Figure 9(a). Running average profiles for reaction stage C. (b). Mean running average profiles for reaction stage C. (c). Mean-adjusted running average profiles for reaction stage C. (d). Average mean-adjusted running average profiles for reaction stage C.

this reaction stage costs around £60 000. Indeed, in this case there is an argument to lengthen the B cycles, at the expense of production, to reduce overall operating costs.

Estimating the average (or total) annual added value of running at a fixed cycle length involves further manipulation of the cycle profiles (similar to those in Figure 7(b) for reaction stage C). A sequence $\{r^{\mathrm{C}}_{ik}\}, k = 1, 2, \dots L^{\mathrm{C}}_i$, termed the running average profile, is defined for the first k days of any individual cycle profile i. This sequence is then calculated for each profile i as

$$r^{\mathrm{C}}_{ik} = \sum_{j=1}^{k} x^{\mathrm{C}}_{ij}/k, \quad i = 1, \dots, N^{\mathrm{C}} \tag{2}$$

Figure 9(a) shows the running average profiles (RAPs) superimposed for reaction stage C. The mean RAP $\{\bar{r}^{\mathrm{C}}_{\cdot k}\}, k = 1, 2, \dots L^{\mathrm{C}}_i$, can then be calculated in a similar manner to the mean cycle profile. This is plotted in Figure 9(b) along with the pointwise 95% confidence band.

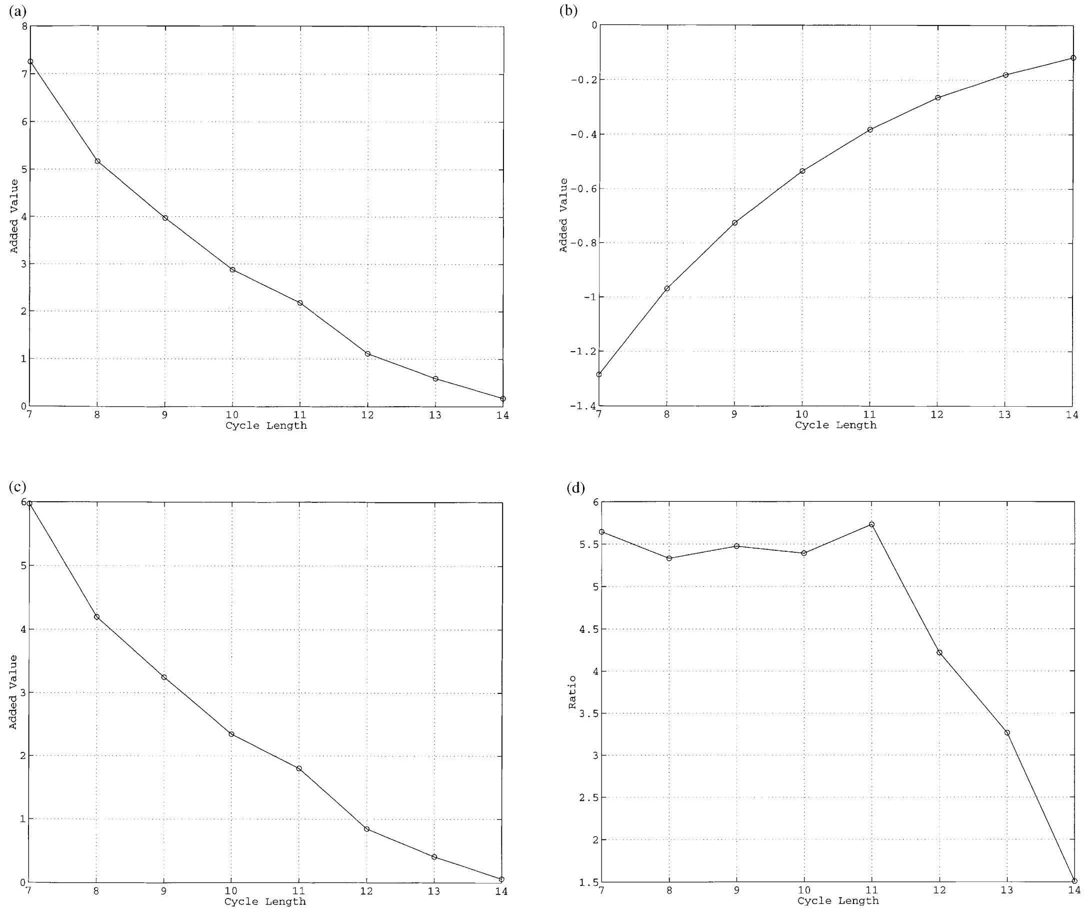

Figure 10(a). Plot of added value from increased production. (b). Annual cost of operating with a fixed cycle length distribution. (c). Plot of net benefit. (d). Absolute ratio of net benefit to added value.

Since interest is in the decay profile rather than the actual starting value of the RAP, each RAP in Figure 9(a) can be adjusted with respect to its mean value over the cycle. This can be achieved without loss of pertinent information. Let $\{\bar{r}^{\mathrm{C}}_{i\cdot}\}, i = 1, 2, \ldots, L^{\mathrm{C}}$, be the mean of each individual cycle i and let $\{\bar{\bar{r}}^{\mathrm{C}}_{\cdot\cdot}\}, i = 1, 2, \ldots, L^{\mathrm{C}}, k = 1, 2, \ldots N^{\mathrm{C}}$, be the overall mean of the individual cycles. The mean-adjusted RAPs are then calculated from

$$Ra^{\mathrm{C}}_{ik} = \{\bar{\bar{r}}^{\mathrm{C}}_{\cdot\cdot} + (r^{\mathrm{C}}_{ik} - \bar{r}^{\mathrm{C}}_{i\cdot})\}, \quad i = 1, 2, \ldots, L^{\mathrm{C}}, \quad k = 1, 2, \ldots, N^{\mathrm{C}} \tag{3}$$

and are shown overlaid in Figure 9(c). The average mean-adjusted RAP $\{\overline{Ra}^{\mathrm{C}}_{\cdot k}\}, k = 1, 2, \ldots, N^{\mathrm{C}}$, in Figure 9(d) exhibits similar trends to the mean RAP $\{\bar{r}^{\mathrm{C}}_{\cdot k}\}$ in Figure 9(b), but the confidence band is narrower.

Using the mean-adjusted RAP, the annual added value of running at a specified maximum cycle length can be calculated. Symbolically, if $\{L^{\mathrm{C}}_{i}\}, i = 1, 2, \ldots, N^{\mathrm{C}}$, are the true cycle lengths, then the added value of running with modified cycles of lengths $\{L_i^{\mathrm{C}*}\}, i = 1, 2, \ldots, N^{\mathrm{C}}$, can be calculated for any particular choice of maximum cycle length $L_{\mathrm{MAX}}^{\mathrm{C}}$:

$$L_i^{\mathrm{C}*} = \begin{cases} -L_i^{\mathrm{C}} & \text{if } L_i^{\mathrm{C}} \leq L_{\mathrm{MAX}}^{\mathrm{C}} \\ L_{\mathrm{MAX}}^{\mathrm{C}} & \text{if } L_i^{\mathrm{C}} > L_{\mathrm{MAX}}^{\mathrm{C}} \end{cases} \quad \text{for } i = 1, 2, \ldots, N^{\mathrm{C}} \tag{4}$$

The curve in Figure 10(a) is an estimate of the added value that would have been achieved had the length of all C cycles larger than a specific cycle length $L_{\mathrm{MAX}}^{\mathrm{C}}$ over the 6 year period been reduced to the specific maximum cycle length. In addition, from the distribution of cycle length, Figure 7(d), the annual cost of operating with a fixed cycle length distribution can be estimated and is shown in Figure 10(b).

The procedure is illustrated through a simple numerical example. Assume that a desired maximum cycle length of 12 days has been selected, i.e. $L_{\mathrm{MAX}}^{\mathrm{C}} = 12$, and of the total number of cycles, four had lengths in excess of $L_{\mathrm{MAX}}^{\mathrm{C}}$, say 15 days. If all the cycles were of length $L_{\mathrm{MAX}}^{\mathrm{C}}$, these four cycles would have resulted in one additional cycle of 12 days, since $(15 - 12) \times 4 = 12$. In this case the added value from increased production would be the product of the value of the mean RAP for day 12 and the market value of the product. The new 12 day cycle would therefore incur an additional cost of one regeneration. Therefore in this particular example the added value from the increased number of cycles would be equal to the cost of one bed regeneration.

The net benefit shown in Figure 10(c) is calculated by summing the values in Figures 10(a) and 10(b) for the respective cycle lengths. The absolute ratio of the net benefit to the added value of increased bed regenerations is the net benefit-to-cost ratio. This is shown in Figure 10(d). The results given in Figures 10(a)–10(d) indicate that operating with a maximum C cycle length of between 9 and 11 days would bring a net annual benefit in the region of £ 250000, taking account of total regeneration costs of around £ 60000. Adoption of this new bed regeneration strategy is therefore economically viable depending on the commercial environment.

A similar cost–benefit analysis applied to the reaction stage is complicated by the fact that interest is in the lengthening, not shortening, of the cycles. Owing to the sparsity of long B cycles available in the data set, it was necessary to either extrapolate or omit a considerable number of bed cycle profiles of relatively short lengths. A number of approaches have been proposed in the literature [16] to achieve this. Results using different types of extrapolation/omission indicate that the current average cycle lengths of 10–12 days are close to optimal.

7. CONCLUSIONS

A range of statistical methods have been used to analyse data from a complex, full-scale petrochemical process, with the objective of optimizing operational cycle lengths in the three major reaction stages. From the initial exploratory analysis, information concerning the operation of the plant was clearly seen. Based on this, a variety of statistical techniques were used to model production rate as a function of process operation, including short- and long-term catalyst decay for the time period following a plant shutdown. Table III summarizes the results from the modelling section. It was found that ridge regression, partial least squares and stepwise selection multiple linear regression, assessed using cross-validation, yielded similar predictive models. It was concluded that production rate decays by approximately 0·15 t h^{-1} day^{-1} in the third reaction stage and by approximately 0·7 t h^{-1} day^{-1} in the second reaction stage. No additional benefit was found from the application of non-linear partial least squares or Curds and Whey.

More extensive comparative studies [15] have shown that there are generally minor differences among the predictive performances of tools where there are more observations than samples, i.e.

Table III. Summary of predictive abilities of different modelling techniques

Method	MSE[a]	Comments
Stepwise multiple linear regression	0·33	
Ridge regression	0·31	Cross-validation was used both to select the shrinkage parameter and estimate the MSE
Linear partial least squares	0·32	Cross-validation was used both to select the number of factors to retain and estimate the MSE
Non-linear partial least squares	0·35	Cross-validation was used both to select the number of factors to retain and estimate the MSE
Curds and Whey	0·32	Generalized cross-validation was used for shrinkage estimation in canonical analysis, and cross-validation was used both to select the common ridge parameter and estimate the MSE

[a] MSE is the mean square error for modelling the standardized response C11=/C12=.

$n \gg p$, as in this problem. In practice, different methods are favoured by practitioners in different fields, e.g. chemometricians have a preference for projection methods such as PLS, whilst statisticians favour more classical regression methods such as MLR and ridge regression. This study has shown that provided the methods are applied and validated appropriately, there are no marked differences between the different approaches.

By treating the times between successive changes of the catalyst in the reactor beds as batches of unequal length, the understanding and isolation of the effects of catalyst decay were enhanced. Once again using the total throughput as the most appropriate engineering variable to describe variable decay, a good estimate of catalyst decay through time was achieved by averaging the total product cycle profiles. The results obtained complimented those derived from the modelling work and were also in good agreement with the understanding of the process personnel. Based on this work, competing economic factors associated with the level of total product and the number of regenerations required to operate at a specific bed cycle, an estimate of the optimum cycle length from an economic perspective was defined. Reduction of cycle lengths in the third reaction stage to approximately 10 days was shown to have substantial economic value.

The paper has comprehensively demonstrated that the application of multivariate data methodologies can lead to enhanced process understanding in complex, real world cases such as the one considered, as well as indicating significant potential economic benefits. It is hoped that in this way the wider use of chemometric techniques will be adopted by the process engineering community working closely with their colleagues skilled in chemometrics and statistics.

ACKNOWLEDGEMENTS

E. Kaskavelis would like to acknowledge the financial and scientific support of the Statistics Group, Shell Research Ltd and the Foresight Centre for Process Analytics and Control Technology, University of Newcastle, UK. The authors would also like to acknowledge the anonomyous referees who have so considerably helped to enhance the paper.

APPENDIX:

NOMENCLATURE

$L^{\mathrm{C}}_{\mathrm{MAX}}$	maximum C bed cycle length
L^{C}_{i}	length of *i*th C bed cycle
n	total number of days
N^{C}	number of C bed cycles
p_i	value of variable for day *i*
r^{C}_{ik}	running average value of *i*th C bed cycle for *k* relative days
t_k	relative days for cycle beds ($k = A$, B, C reactors)
$\bar{r}^{\mathrm{C}}_{i\cdot}$	mean of *i*th running average profile for C beds
$\bar{r}^{\mathrm{C}}_{\cdot k}$	mean running average profile for *k* relative days for C beds
$\bar{\bar{r}}^{\mathrm{C}}_{\cdot\cdot}$	overall mean of running average values
Ra^{C}_{ik}	'mean-adjusted' running average profile for C cycles
$\overline{Ra}^{\mathrm{C}}_{\cdot k}$	mean of 'mean-adjusted' running average profiles for C cycles
$t_{\mathrm{A,B,C}}$	relative days for A, B and C cycles
t	day number for long-term catalyst effect
T^{C}_{i}	start day of *i*th cycle for C beds
x^{C}_{ij}	*j*th relative day of *i*th cycle
$\bar{x}^{\mathrm{C}}_{\cdot j}$	mean value for *j*th relative day of C bed cycles

REFERENCES

1. Draper NR, Smith H. *Applied Regression Analysis*. Wiley: New York, 1998.
2. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **42**: 80–86.
3. Geladi P, Kowalski BR. Partial least squares regression: a tutorial. *Anal. Chim. Acta* 1986; **185**: 1–17.
4. Breiman L, Friedman JH. Predicting multivariate responses in multiple linear regression (with discussion). *J. R. Statist. Soc. B.* 1997; **59**: 3–54.
5. Wold S, Kettaneh-Wold N, Skagerberg B. Non-linear PLS modelling. *Chemometrics Intele. Lab. Syst.* 1989; **7**: 53–65.
6. Baffi G, Martin EB, Morris AJ. Non-linear projection to latent structures revisited (the quadratic PLS algorithm). *Comput. Chem. Engng* 1999; **23**: 395–411.
7. Malthouse EC, Tamhane AC, Mah RSH. Non-linear partial least squares. *Comput. Chem. Engng* 1997; **21**: 875–890.
8. Qin SJ, McAvoy TJ. Non-linear PLS modelling using neural networks. *Comput. Chem. Engng* 1992; **16**: 379–391.
9. Wilson DJH, Irwin GW, Lightbody G. Non-linear PLS modelling using radial basis functions. *Proc. Am. Control Conf.*, Albuquerque, NM, 1997.
10. Nelson PRC, Taylor PA, MacGregor JF. Missing data methods in PCA and PLS: score calculations with incomplete observations. *Chemometrics Intell. Lab. Syst.* 1996; **35**: 45–65.
11. Jackson JE. *A User's Guide to Principal Components*. Wiley: New York, 1991.
12. Stone M. Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc. B* 1974; **36**: 111–147.
13. Jonathan P, McCarthy WV, Krzanowski WJ. On the use of cross-validation to assess performance in multivariate prediction. *Statist. Comput.* 2000; **10**: 209–229.
14. S-PLUS software. Available: http://www.mathsoft.com/splus.
15. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics* 1993; **35**: 109–148.
16. Rothwell SG, Martin EB, Morris AJ. Comparison of methods for dealing with uneven length batches with application to MPLS prediction of batch bioprocess quality. Proc. IFAC DYCOPS-5, 1998; 66–71.