TITLE:

Integration of Stochastic Simulation with Multivariate Analysis: Short Term Facility
Fit Prediction

AUTHORS:

Adam Stonier^{a,b}, David Pain^b, Ashley Westlake^b, Nicholas Hutchinson^c, Nina F.
Thornhill^d, Suzanne S. Farid^a.

^a The Advanced Centre for Biochemical Engineering, Department of Biochemical
Engineering, University College London, Torrington Place, London WC1E 7JE, UK

^b Lonza Biologics plc., 228 Bath Road, Slough, SL14DX, UK

^c Parker Hannifin Ltd, Parker domnick hunter, Process Division, Durham Road,
Durham, DH3 2SF, UK

^d Centre for Process Systems Engineering, Department of Chemical Engineering,
Imperial College London, South Kensington Campus, London SW7 2AZ, UK

Corresponding author:

Dr Suzanne S. Farid

University College London, Department of Biochemical Engineering, Torrington
Place, London, WC1E 7JE. UK

s.farid@ucl.ac.uk

+44 20 7679 4415

ABSTRACT:

This paper describes a decision-support tool that integrates Monte Carlo simulation data derived using a stochastic discrete-event simulation model to mimic process fluctuations with advanced multivariate statistical techniques to help pinpoint the potential root causes of sub-optimal short term facility fit issues. Principal component analysis combined with clustering algorithms was used to analyse the complex datasets from complete industrial batch processes for biopharmaceuticals. The challenge of visualising the multidimensional nature of the dataset was addressed using hierarchical and K-means clustering as well as parallel co-ordinate plots to help identify process fingerprints and characteristics of clusters leading to sub-optimal facility fit issues. Industrially-relevant case studies are presented that focus on technology transfer challenges for therapeutic antibodies moving from early phase to late phase clinical trials. The case study details how sub-optimal facility fit can be alleviated by allocating alternative product pool tanks. The impact of this operational change is then assessed by reviewing an updated process fingerprint.

KEYWORDS:

biopharmaceutical manufacturing processes, stochastic decision-support tools, discrete-event simulation, multivariate analysis, principal component analysis, parallel co-ordinate plots

1. Introduction

Long term facility fit issues are generally concerned with the future proofing of manufacturing facilities and platform processes (Stonier et al., 2009, 2012). Any challenges highlighted can be addressed over time with the development of new technologies and large capital projects. Short term facility fit issues are those that arise due to the batch-to-batch variability associated with a single process and there may be insufficient time to make significant changes to facilities so early identification of issues is crucial.

Tech transfer of pilot scale processes into facilities during late phase clinical trials of a drug candidate can often lead to short term facility fit issues (e.g. Kelley, 2009; Chang, 2011). The impact of process fluctuations becomes more pronounced during tech transfer since mismatches in equipment sizes are not easy to adapt to at large scale. Predicting and understanding the root causes of sub-optimal facility fit becomes critical so as to minimise the probability of having to discard expensive product and remove barriers to robust, scalable process design. The ideal is for a process in development to complete technology transfer to full scale manufacturing with no redevelopment costs or surprises. Essential to achieving this is a systematic method for analysing large complex datasets and extracting critical combinations of fluctuations that lead to product loss and scheduling delays. This paper explores how best to integrate large datasets, derived from stochastic simulations of tech transfer batch fluctuations, with advanced multivariate analysis and visualisation techniques such as clustering, principal component analysis (PCA), and parallel co-ordinate plots so as to help identify the potential root causes of sub-optimal short term facility fit issues.

Multivariate analysis techniques such as PCA have been used in a range of biotechnology applications such as analysing vaccine production runs to identify rejected batches and sources of variation (Thomassen et al, 2010), understanding the impact of raw material variability on cell culture performance (Lee et al, 2012; Kirdar et al, 2010), analysing the impact of chromatography operating conditions on chromatogram profiles (Pate et al., 1999, 2004; Malmquist and Danielsson, 1994; Larson et al., 2003; Edwards-Parton et al., 2008; Hou et al., 2011), and assessing process comparability for cell culture operations across different scales of production (Kirdar et al, 2007).

Most works on MVA utilise historical datasets. However, the nature of tech transfer activities where a single batch is usually run to identify facility fit issues and fix them before further batches are operated means that such datasets do not exist for this problem. However, companies with experience in tech transfer activities can identify typical process fluctuations seen from the expected base case performance. However, their combined impact is not known nor their contribution to product loss. Hence in this work Monte Carlo simulations were used to generate the range and likelihood of possible outcomes in terms of process variation and the consequences on key metrics such as product mass loss and cost of goods (COG).

This paper builds on the flexible database-driven simulation platform described in Stonier et al, (2009, 2012) that captures the mass balances, equipment sizing, dynamic resource allocation and process economics of purification sequences in monoclonal antibody manufacturing processes. Previous work by the authors (Stonier et al., 2009, 2012) discusses the use of the tool to perform a small number of simulation runs to generate sufficient data to assess the impact of a single parameter change. This paper describes the extension of the tool to mimic the stochastic nature

of industrial batch processes when transferred to large-scale facilities and to identify the potential root causes of short term facility fit issues. This was achieved by building in capabilities to run Monte Carlo simulations and exploring how best to integrate stochastic results with advanced multivariate statistical analysis techniques.

Monte Carlo simulations result in very large datasets and hence visualising the results of the MVA becomes more challenging. Although PCA might reduce hundreds of datasets to a few principal components, it does not automatically identify clusters of batches with similar characteristics for further examination. Hence in this paper, algorithms adapted from Thornhill et al (2006) are used to achieve hierarchical and k-means clustering of the datasets so as to identify significant clusters of batches. Multidimensional visualization of each cluster's characteristics in terms of the raw data (e.g. cell culture titre) is achieved through the generation of novel multiple 'stacked' parallel co-ordinate plots; this is in contrast to other works (eg Wang et al., 2004) where parallel co-ordinate plots are used to plot PC scores rather than reverting to the actual process data. The resulting facility fingerprints enhance process understanding of interactions between variables in an informative and clear manner.

2. Problem Domain

Upon technology transfer of a process from pilot scale to large scale, a GMP engineering batch is normally carried out to test performance and identify facility fit issues. Until this point, data is only available from the laboratory and pilot scales of the process. The first batch at large scale is subject to a much greater degree of uncertainty that would normally be expected between subsequent batches. This uncertainty is due in part to scale effects such as variability introduced by operating larger chromatography columns or increased product holdup in systems. The impact

of this variability becomes more exaggerated at large scale where any equipment limitations are difficult to adapt to. As a result, the first batch at large scale is more difficult to predict than any other batch. Four key parameters were identified as key to the variability at scale: product titre, step yields, chromatography eluate volumes and filter flux rates. Representative triangular distributions were derived through discussion with industrial experts (Table 1). These four sources of variability are discussed further below.

2.1. Product Titre

This variation in the total amount product entering into the purification process is felt across all chromatography steps since the number of cycles is determined by the mass load of product, making it an important parameter to capture. Some variation in average product titres upon scale up is expected due to the inherent uncertainty associated with biological systems and whilst not fully understood, contributing factors include: changes in vessel geometry and subsequent hydrodynamic affects, the differing levels of control of critical factors such as pH, temperature and gas mixtures within different reactors. The impact of these factors is complex and results in varying levels of uncertainty for different products and cell lines. The value of $\pm 10\%$ was determined through discussions with industrial advisors and represents a reasonable level of variation for a well-understood process from pilot scale to large scale manufacturing. The degree with which titres may vary however can be difficult to predict and a wider range may need to be considered for some products, especially where the differences in scale upon tech transfer are large.

2.2. Step Yields

Variation in yield losses across steps are typically small as represented by the relatively narrow ranges specified in Table 1. The case studies in this paper are concerned with a hypothetical first batch at scale where fine tuning with respect to minimising yield losses such as optimising flush volumes to minimise losses through holdup may not have been fully identified. Hence this variation may be reduced further over subsequent batches. A larger range specified for chromatography operations is related to issues surrounding eluate volumes as discussed below.

2.3. Chromatography Eluate Volumes

Variability in the eluate volumes from the chromatography operations was highlighted as one of the largest factors impacting facility fit at scale. When using collection criteria based on UV traces, a small variation in the position of the peak can have a large effect on the volume collected particularly for steps where significant leading or tailing on elution peaks may occur. Certain chromatography steps for example may be highly sensitive to the pH and conductivity of the elution buffer. When considering a process at scale which may use inline dilution for buffers, the control over pH and conductivity is generally less precise than can be achieved at the laboratory scale. The result may be different binding profiles. Different batches of buffers can also influence the binding and elution profiles. Furthermore, when tech transfer activities are planned prior to the process limits evaluation validation studies during Phase III clinical trials, most of the manufacturing data would be from lab-scale or small-scale experiments with little indication of the potential variability in elution performance. As a result, a value of $\pm 50\%$ in eluate volumes was considered to reflect this uncertainty as a worst case scenario for most processes.

2.4. Filter Flux Rates

Filter flux rate variability can impact on the process schedule. The different pressure and flow rate capabilities of equipment at different scales can lead to uncertainty in the achievable flux rates. These are typically in the $\pm 10\%$ range.

3. Method

A database-driven discrete-event simulation tool (Stonier et al. 2009 and submitted) for modelling the logistics and process economics of antibody process was used as the core evaluation engine for this study. It was adapted to perform Monte Carlo simulations and handle the larger datasets required so as to mimic the impact of process fluctuations on the key outputs. A series of multivariate and visualisation analyses were then explored to examine the stochastic simulation outputs. These are elaborated upon further in the following sections.

3.1. Tool Overview

A database-driven discrete-event simulation tool was built to model the logistics and economics of alternative purification strategies. The overall structure along with the key input and output parameters is shown in Stonier et al. (2012). The benefits of a database-driven approach using MySQL (MySQL AB, Uppsala, Sweden) were harnessed to capture the process, business and risk features of multiple biopharmaceutical purification sequences in a multiproduct facility and better manage the large datasets required for multiple processes, uncertainty analysis and optimisation. The dynamic features of bioprocesses and impact of resource constraints, including suite availability, were captured through linkage to a discrete-event simulation engine, Extend (Imagine That! Inc, San Jose, USA) that ran in the

background requiring no end-user interaction. This was achieved by adopting a router-based approach to automatically adapt to a number of process and facility configurations that may be specified by the user in the database without the need for recoding. Universal modelling language (UML) was used to provide a conceptual representation of the hierarchical breakdown of key activities and facility resources used in purification campaigns along with the relationships between them. The UML hierarchical structure was mapped to the database where it was used to define the data tables and relationships and to the simulation engine where it defined the key procedures to be implemented.

3.2. Database Configuration for the Monte Carlo Simulations

The simulation database was modified to handle the more complex datasets from stochastic simulations. In this paper only triangular distributions were used, but the new data structure was designed to allow different distribution functions to be defined in the future. The distributions were defined with two fields in the database tables. The fields were defined in addition to the existing data so as to allow the simulation to be operated in deterministic mode without modification. The first field was a numerical identifier which relates to the shape of the distribution. The second field was a coded variable containing the parameters for the bounds of the distribution. The value in the identifier field determined which function would be used by the simulation and the coded variable stored the inputs to the function as a set of comma separated variables. This approach to capturing stochastic inputs enabled modifications to be made that had limited effect on the existing data structure whilst allowing for the maximum degree of flexibility.

As well as defining a data structure for the input parameters, the results from multiple iterations were stored in such a way so as to allow the key inputs and outputs of each iteration to be isolated for more in-depth analysis on the root causes of the output observed.

3.3 Stochastic Discrete-Event Simulation Engine

The discrete-event simulation engine described in Stonier et al. (2009, 2012) was configured to run Monte Carlo simulations by tying a random number generator into new functions which could generate the values within the distributions described in the database. The numerical identifier field was used to determine if a distribution should be used. A value of 1 directs the simulation engine to use the original deterministic (point) value. Values other than 1 correspond to different shapes of distribution (e.g. a triangular distribution).

3.4. Principal Component Analysis Method

Methods for principal component analysis have been widely documented. The method used in this paper was derived from Thornhill et al. (2006) and uses singular value decomposition. The output data from the simulations were first transformed by mean centring and stored as a comma separated variable file which was then imported. The PCA analysis was carried out using SPSS Statistics (IBM Corp., NY).

The main aim of principal component analysis is to reduce the dimensionality of a dataset whilst minimising the loss of information. The primary output of the analysis is a new data set where the original variables have been replaced by principal components and the values for each variable are transformed into scores within each principal component. The weighting of the original variables within the principal

components will determine how the scores relate back to the original data. However, it can be said that the structure of the scores will represent the structure of the original dataset and if successful the structure will be evident in a lower dimension than the original data and thus easier to visualise.

Eigenvalues are used to determine which principal components are of key importance. Several methods are used to test for significance. The most common of these is to divide each eigenvalue by the mean of all eigenvalues and consider all principal components with an eigenvalue greater than 1 as significant. The eigenvalues can also be plotted for each principal component, the result is a scree plot as shown in Figure 5. Visual analysis of the scree plot may determine that the average eigenvalue method is not suitable if the data is not highly correlated. Instead a visual analysis of the scree is required to identify significant changes in gradient. The ideal scree plot will consist of a small number of high value eigenvalues followed by a sudden drop and a gradual tailing of the remaining eigenvalues. The principal components before the drop are considered significant and those after have less statistical value.

Upon identification of the key principal components the scores from each component can be analysed using cluster analysis. SPSS was used to assess the performance of two clustering methods. These were hierarchical clustering, and K-means clustering. Hierarchical clustering was used by Thornhill et al. (2006) with a data set of similar complexity and structure and whilst various methods were used to improve the method, hierarchical clustering was not able to satisfactorily elucidate the clusters within the dataset generated in this analysis. An example dendrogram is shown in Appendix A. Instead K-means clustering was used. The K-means clustering algorithm used in this paper was a function within SPSS.

SPSS was also used to generate graphical outputs in the form of scatter plot matrices and parallel co-ordinate plots. Both methods were used to visualise the higher dimensional datasets generated as part of this analysis and are discussed in more detail throughout this paper.

3.5. Setup

The case study looks at the first batch at scale. The process description for a hypothetical 10,000L facility is summarised in Fig 1 and Table 2. The aim was to identify potential short term facility fit issues upon tech transfer and the probability that they would be realised. In addition to facility fit issues, the case study also highlights how the sensitivity of a process to uncertainty can change with respect to facility design. The simulation engine was configured to run Monte Carlo simulations. The results were first analysed to investigate potential facility fit issues. Modifications were then made to the simulated process to improve facility fit and the Monte Carlo simulations were repeated with the improved process. The complexity of each dataset was then reduced using principle component analysis combined with clustering to reduce the dimensionality and eliminate noise (Thornhill et al. 2006). The results were compared to assess the impact of the process change.

The appropriate number of principal components was selected by a visual analysis of the eigenvalues plotted for each principal component. After selection of the principal components, clusters in the dataset were identified using K-means clustering. The clusters were generated by analysing the component scores. Finally, the input parameter distributions for each cluster was visualised on vertical parallel co-ordinate plots. The Monte Carlo simulation was run for 1000 iterations to generate the following results.

4. Results

4.1 Identifying Sub-optimal Facility Fit

4.1.1 Mass Throughput Profile

One of the key parameters of interest is the mass throughput of the facility (kg/batch). In many cases the planning of manufacturing campaigns will be determined well in advance of the first batches at scale, therefore it is important to have an accurate prediction of facility throughput early on in the development lifecycle. The stochastic tool was used to predict the likelihood of product loss upon tech transfer using Monte Carlo simulation, given the expected fluctuations in key performance indicators and purification operating parameters indicated in Table 1. Figure 2 shows the predicted facility throughput for the processes running in the 10,000L facility. Based on deterministic values, the predicted facility throughput would be calculated at 11Kg/batch. The values predicted by the simulation fall well short of this value. A very small proportion of batches meet the expected throughput. This is suggestive of facility fit issues and prompted further investigation.

4.1.2. Risk Hotspots for Facility Fit

The Monte Carlo simulation data for the pool volumes after each step in the purification process was examined to identify the location of the equipment limitations causing the facility fit issues. The simulation engine maintained a log of events that occurred during each simulation run. A section of this log was dedicated to vessel operations and an error event was recorded when the amount of material exceeded a tank's maximum volume. Review of the log showed that an over-fill event occurred in two of the product pool tanks in a number of runs. Since step yields and

eluate volumes are stochastic variables some variability of pool volumes was expected. Figure 3 shows the distribution of the pool volumes at each step. With the exception of the harvest pool which in this simulation is defined as a fixed value entering the process, variation in the pool volumes is evident. Figure 3(a) identifies the virus inactivation and anion exchange chromatography pool volumes as risk hotspots.

The vertical histogram plotted alongside each column of data points shows a spike in the distributions at 2500L for both of these tanks. This is due to the fact that the largest volume that can be stored in these tanks was 2500L. Surplus volume was diverted to waste and the product was lost impacting throughput on a large number of batches. This is a facility fit issue which must be addressed. In this case a facility modification was possible to re-route the product into alternative larger tanks. The effect of this modification is shown in Figure 3(b)

With the tank volumes increased to handle the predicted pools, the volume of the spike in the distribution was removed. The result of this modification would be to increase the volume entering into subsequent steps and could also move the bottleneck downstream. In this instance however sufficient capacity was installed and no additional tank issues were apparent. The impact of this facility fix on the facility throughput can be seen in Figure 2 (dotted line), with the throughput closer to the predicted value of 11kg/batch.

4.1.3. Probability of Meeting Demand

The data shown in Figure 4 can be used to generate probabilities of being able to manufacture sufficient material within a number of batches and hence better manage manufacturing risk. Figure 4 shows the probability of being able to meet varying demands within three batches. This illustrates that the minimum guaranteed demand

for the modified facility is almost double that of the base case facility. This is because not only is the throughput of the modified facility higher, it is less variable.

4.2 Principal Component Analysis

In order to gain more insight into the performance of the process in the 10,000L facility principal component analysis was used to reduce the complexity of the dataset. The two datasets from the base case and modified facility were analysed individually using the same procedure. Firstly the number of significant principal components was identified by plotting their eigenvalues as shown in Figure 5.

4.2.1 Principal Component Selection

The average eigenvalue method, which defines significant components as having eigenvalues greater than 1 is not suitable for these datasets as this would include a number of principal components which add little value. Instead the appropriate number of principal components was selected by a visual analysis of the dataset. In this case for the base case the first 5 principal components were selected and for the modified facility the first 6. In both cases selection of any additional components added little to the integrity of the dataset.

4.2.2 Cluster Identification

After selection of the principal components, hierarchical clustering was used to determine the number of significant clusters by visual inspection of the resulting dendrogram (Appendix 1) which clearly shows three parent clusters dividing into six distinct clusters (at an average distance between clusters of 20 units). Further branching leads to less distinct clusters. The method of k-means clustering was then used to identify the membership of six distinct clusters. To aid visualisation, the key

principal components were plotted against each other on a scatter plot matrix (Figure 5). Each of the cells in the scatter plot matrices shown in Figure 5 were generated by plotting the component scores from each principal component against each other. For example, all the charts in the first row show principal component 1 scores plotted against: themselves in column 1, principal component 2 scores in column 2 and principal component 3 scores in column 3. Since the plots on the diagonal are the principal components plotted against themselves the data is represented as a histogram to show the distribution of the scores within each principal component.

In addition to the scatter plot matrices the clusters identified by k-means clustering are shown as different coloured points. For example, the scatter plots in Fig 6a for PC3 v PC1 clearly indicate the presence of two distinct and non-overlapping clusters. Each of these clusters can be split into further non-overlapping clusters as indicated in Fig 6a as a result of k-means clustering in higher dimensions. By allowing the clusters to be visualised in the PCA score space, these scatter plots allow verification of the success of the k-means clustering for these datasets. By carrying out the principal component analysis first, the parameters which have less influence on the structures within the dataset have been removed thereby reducing noise and the dimensionality of the data.

4.2.3. Parallel Co-ordinate Visualisation

Next, the clustered data was converted back to values which have more meaning. This was done by separating the input dataset into the clusters identified by the k-means algorithm and this time plotting the original variables instead of the component scores on a series of 'stacked' parallel co-ordinate plots. This data is shown in Figure 7. Figure 7(a) shows the data from the base case 10,000L facility. Thirty parameters were input into the PCA algorithm however only the parameters of interest are shown

here. For clarity each parameter is normalised between -1 and 1 such that they can be easily compared. The result is a facility fingerprint. By comparing Figures 7(a) and (b) the wider impact of changing the tank volume can be quickly assessed.

Each cluster of batches in Figure 7 has common characteristics that combined lead to different outcomes such as high mass losses. For example, an examination of cluster 4 in Figure 7(a) indicates that despite high titres, the cluster also exhibits high cost of goods. This can be attributed to a high average level of mass loss that was strongly linked to a high average eluate volume on the affinity step.

Figures 7(a) and (b) can be compared together to understand how the facility fingerprint changes after the facility fix. Before the fix (Figure 7(a)), all clusters of batches exhibit significant variance in product mass loss. After the fix (Figure 7(b)), the variance in mass loss can be seen to be removed as indicated by the pinch point formed across all clusters. Further comparison links this to the reduced variability in the number of chromatography cycles (affinity, AEX and CEX). This represents a significant improvement in the process schedule's ability to absorb the uncertainty. The number of cycles overall may be higher in the modified facility, however they are more predictable which is often of greater importance.

5. Conclusions

This paper illustrates the insights that can be gained by integrating stochastic simulation data with advanced multivariate statistical techniques for predicting and understanding facility fit issues upon tech transfer. The Monte Carlo data enabled a greater degree of understanding of the impact of process fluctuations and subsequent facility fixes on the likelihood of meeting manufacturing targets. An early appreciation of risk allows for the planning of risk avoidance strategies such as

contingency batches or process changes. The effect of the facility fit identified was analysed in more detail with multivariate statistical analysis techniques that harnessed principal component analysis, clustering algorithms and high dimensional visualisation techniques combined to generate facility fingerprints. These allowed for rapid identification of process robustness and characteristics of clusters of batches that resulted in product losses and hence sub-optimal facility fit.

Acknowledgements

Financial support from the UK Engineering and Physical Sciences Research Council (EPSRC) and Lonza Biologics is gratefully acknowledged. Support from the EPSRC Centre for Innovative Manufacturing in Emergent Macromolecular Therapies and its partners is also acknowledged.

References

- Chang J. Process Validation Challenges for Tech Transfer. *Pharmaceutical Outsourcing* 2011; 12 (4).
- Edwards-Parton S, Thornhill NF, Bracewell DG, Liddell JM, Titchener-Hooker NJ. Principal component score modeling for the rapid description of chromatographic separations. *Biotechnol Prog* 2008; 24(1): 202–208.
- Hou Y, Jiang C, Shukla AA, Cramer SM. Improved process analytical technology for protein A chromatography using predictive principal component analysis tools. *Biotechnol Bioeng* 2011; 108: 59-68.
- Kelley B. Industrialization of mAb production technology The bioprocessing industry at a crossroads. *Mabs* 2009; 1(5):443-452.
- Kirdar AI, Chen G, Weidner J, Rathore AS. Application of nearinfrared (NIR) spectroscopy for screening of raw materials used in the cell culture medium for the production of a recombinant therapeutic protein. *Biotechnol Prog.* 2010;26:527–531.
- Kirdar AO, Chen G, Rathore S. Combining Near-Infrared (NIR) Spectroscopy and Multivariate Data Analysis (MVDA) for Screening of Raw Materials Used in the Cell Culture Medium for the Production of a Recombinant Therapeutic Protein, *Biotechnol. Prog.* 2010; 26: 527-531.
- Kirdar AO, Conner JS, Baclaski J, Rathore AS. Application of multivariate analysis toward biotech processes: case study of a cell-culture unit operation. *Biotechnol Prog.* 2007;23:61–67.

- Larson TM, Davis J, Lam H, Cacia J. Use of process data to assess chromatographic performance in production-scale protein purification columns. *Biotechnol Prog* 2003; 19(2):485–492.
- Lee HW, Christie A, Liu JJ, Yoon S. Estimation of raw material performance in mammalian cell culture using near infrared spectra combined with chemometrics approaches. *Biotechnol Prog* 2012; 28: 824–832.
- Malmquist G, Danielsson R. Alignment of chromatographic profiles for principal component analysis—A prerequisite for fingerprinting methods. *J Chromatogr A* 1994; 687(1):71–88.
- Pate ME, Turner MK, Thornhill NF, Titchener-Hooker NJ. Principal component analysis of nonlinear chromatography. *Biotechnol Prog* 2004;20(1):215–222.
- Pate ME, Turner MK, Thornhill NF, Titchener-Hooker NJ. The use of principal component analysis for the modelling of high performance liquid chromatography. *Bioprocess Eng* 1999; 21(3):261–272.
- Stonier A, Simaria AS, Smith M, Farid SS. Decisional tool to assess current and future process robustness in an antibody purification facility. *Biotechnol Prog* 2012 (online).
- Stonier A, Smith M, Hutchinson N, Farid SS. Dynamic simulation framework for design of lean biopharmaceutical manufacturing operations. *Computer-Aided Chemical Engineering* 2009; 26: 1069-1074.
- Thomassen YE, van Sprang ENM, van der Pol LA, Bakker WAM. Multivariate data analysis on historical IPV production data for better process understanding and future improvements. *Biotechnol Bioeng*. 2010;107:96–104
- Thornhill NF, Melbø H, Wiik J. Multi-dimensional visualization and clustering of historical process data. *Ind Eng Chem Res* 2006; 45(17): 5971-5985.
- Wang XZ, Medasani S, Marhoon F, Albazzaz H. Multidimensional visualization of principal component scores for process historical data analysis. *Ind. Eng. Chem. Res.* 2004; 43: 7036-7048.

Tables

Table 1: Variable Distribution Ranges

Variable	Min	Most Likely	Max
Product Titre	-10%	Base Case	+10%
Eluate Volumes	-50%	Base Case	+50%
Filter Flux Rates	-10%	Base Case	+10%
Step Yields			
Chromatography Steps	83%	95%	93%
Virus Inactivation	98%	99%	100%
Ultrafiltration/Diafiltration	90%	95%	99%
Virus Retention Filtration	90%	95%	99%

Table 2: Facility Specification at 10,000L scale and expected 2g/L titre

CHROMATOGRAPHY			
	Protein A	AEX	CEX
Column diameter (m)	1	1	1
Bed height (m)	0.20	0.25	0.15
Bed volume (L)	157	196	118
Load capacity (g/L)	25	50	15
Linear velocity (cm/h)	450	450	140
Expected number of cycles	6	2	8
Expected pool* volume (CV/cycle)	2	5.5	3
Expected step yield (%)	88	88	88
Virus inactivation volume (ml/L)	58	N/A	N/A
FILTRATION			
	Post Protein A UF/DF	Post AEX UF	Final UFDF
Retentate tank volume (L)	2500	2500	5000
Expected ave. flux for UF/DF(L/m²h)	100/55	110/60	140/80
Target concentration (g/L)	25	25	38
Diafiltration volumes	3	0	10
Expected step yield (%)	99	99	99

Note: Pool volume refers to the volume of the product stream. In Protein A and CEX steps operated in bind-and-elute mode this refers to the eluate volume collected. In AEX operated in flow-through mode this refers to the load volume plus a proportion of the post wash volume collected.

List of figures

Figure 1: Base case process and facility. On the left the process sequence is shown along with estimated processing times. Note the process is split into three sections with each section aligned to a class of suite. The centre of the diagram shows the number and type of processing suites within the facility. The circles represent the different suites with the arrows showing the route process material can take. CC = cell culture, Cent = centrifugation, ProA = Protein A, VI = virus inactivation, UFDF = Ultrafiltration/Diafiltration, AEX = anion exchange, CEX = cation exchange, VRF = virus retention filtration, DSP = downstream processing, VS = virus secure.

Figure 2: Probability distributions for the throughput predicted from both the base case (-----) and modified (- - -) 10,000L facility. Fluctuations assumed for cell culture titre ($2\text{g/L} \pm 0.2$; harvest $\text{kg/batch} = 20 \pm 2$), step yields, eluate volumes, filtration flux rates. (n=1000)

Figure 3: Distribution of pool volumes for each step given expected fluctuations in the (a) base case and (b) modified 10,000L facility data. Vertical bars are composed of a series of horizontal dashes, each representing the pool volume from a single simulation iteration. For clarity the distribution of values is shown by the histogram alongside each bar. Fluctuations assumed for cell culture titre ($2\text{g/L} \pm 0.2$; harvest $\text{kg/batch} = 20 \pm 2$), step yields, eluate volumes, filtration flux rates. Note: Pool volume refers to the volume of the product stream. In Protein A and CEX steps operated in bind-and-elute mode this refers to the eluate volume collected. In AEX operated in flow-through mode this refers to the load volume plus a proportion of the post wash volume collected. (n=1000).

Figure 4: Probability plot showing the probability of meeting a range of demands within three batches using either the base case (-----) and modified (- - -) 10,000L facility. Fluctuations assumed for cell culture titre ($2\text{g/L} \pm 0.2$; harvest $\text{kg/batch} = 20 \pm 2$), step yields, eluate volumes, filtration flux rates. (n=1000)

Figure 5: Scree plots generated from the principal component analysis of the (a) base case and (b) modified 10,000L facility data.

Figure 6. Clustered scatter matrices showing the score values for the first three principal components generated by the PCA of the (a) base case and (b) modified 10,000L facility data. k-means clustering was used on the first 5 principal components, 3 are shown for clarity. Along the diagonal, histograms are shown of the distributions of scores within each principal component, (n=1000)

Figure 7: Vertical parallel co-ordinate plot showing PCA clustered data for the (a) base case and (b) modified 10,000L facility data. Fluctuations assumed for cell culture titre ($2\text{g/L} \pm 0.2$; harvest $\text{kg/batch} = 20 \pm 2$), step yields, eluate volumes, filtration flux rates. Note: for AEX operated in flow-through mode, the AEX eluate vol refers to the load volume plus a proportion of the post wash volume collected. (n=1000)

Figure A.1: An example dendrogram generated from the outputs from principal component analysis of the base case 10,000L facility data used to identify 6 distinct clusters (seen where distance between clusters = 20 units).

.

Figure 2: Probability distributions for the throughput predicted from both the base case (-----) and modified (- - -) 10,000L facility. Fluctuations assumed for cell culture titre ($2\text{g/L} \pm 0.2$; harvest $\text{kg/batch} = 20 \pm 2$), step yields, eluate volumes, filtration flux rates. ($n=1000$)

Figure 3: Distribution of pool volumes for each step given expected fluctuations in the (a) base case and (b) modified 10,000L facility data. Vertical bars are composed of a series of horizontal dashes, each representing the pool volume from a single simulation iteration. For clarity the distribution of values is shown by the histogram alongside each bar. Fluctuations assumed for cell culture titre ($2\text{g/L} \pm 0.2$; harvest $\text{kg/batch} = 20 \pm 2$), step yields, eluate volumes, filtration flux rates. Note: Pool volume refers to the volume of the product stream. In Protein A and CEX steps operated in bind-and-elute mode this refers to the eluate volume collected. In AEX operated in flow-through mode this refers

to the load volume plus a proportion of the post wash volume collected. (n=1000).

Figure 4: Probability plot showing the probability of meeting a range of demands within three batches using either the base case (-----) and modified (- . -) 10,000L facility. Fluctuations assumed for cell culture titre ($2\text{g/L} \pm 0.2$; harvest $\text{kg/batch} = 20 \pm 2$), step yields, eluate volumes, filtration flux rates. ($n=1000$)

(a)

(b)

Figure 5: Scree plots generated from the principal component analysis of the (a) base case and (b) modified 10,000L facility data.

Figure 6. Clustered scatter matrices showing the score values for the first three principal components generated by the PCA of the (a) base case and (b) modified 10,000L facility data. k-means clustering was used on the first 5 principal components, 3 are shown for clarity. Along the diagonal, histograms are shown of the distributions of scores within each principal component, (n=1000)

(a)

(b)

Figure 7: Vertical parallel co-ordinate plot showing PCA clustered data for the (a) base case and (b) modified 10,000L facility data. Fluctuations assumed for cell culture titre ($2\text{g/L} \pm 0.2$; harvest $\text{kg/batch} = 20 \pm 2$), step yields, eluate volumes, filtration flux rates. Note: for AEX operated in flow-through mode, the AEX eluate vol refers to the load volume plus a proportion of the post wash volume collected. (n=1000)

Appendix 1

Figure A.1: An example dendrogram generated from the outputs from principal component analysis of the base case 10,000L facility data used to identify 6 distinct clusters (seen where distance between clusters = 20 units).