Computers and Chemical Engineering 132 (2019) 106620

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Multivariate statistical process control of an industrial-scale fed-batch simulator

Carlos A. Duran-Villalobos[a,*], Stephen Goldrick[b], Barry Lennox[a]

[a] School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, UK
[b] Department of Biochemical Engineering, University College London, London WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 25 April 2019
Revised 6 October 2019
Accepted 21 October 2019
Available online 24 October 2019

Keywords:
Optimal control
Batch to batch optimisation
Model predictive control
Data-driven modelling
Missing data methods
Partial least square regression

ABSTRACT

This article presents an improved batch-to-batch optimisation technique that is shown to be able to bring the yield closer to its set-point from one batch to the next. In addition, an innovative Model Predictive Control technique is proposed that over multiple batches, reduces the variability in yield that occurs as a result of random variations in raw material properties and in-batch process fluctuations. The proposed controller uses validity constraints to restrict the decisional space to that described by the identification dataset that was used to develop an adaptive multi-way partial least squares model of the process. A further contribution of this article is the formulation of a bootstrap calculation to determine confidence intervals within the hard constraints imposed on model validity. The proposed control strategy was applied to a realistic industrial-scale fed-batch penicillin simulator, where its performance was demonstrated to provide improved consistency and yield when compared with nominal operation.

1. Introduction

In the pharmaceutical industry, regulatory authorities such as the Food and Drugs Agency (FDA) are encouraging the adoption of Quality by Design (QbD) enabling improved product quality through enhanced process control (Yu et al., 2014). Optimal control strategies can maximize product quality by reducing product variability and defects.

To implement these strategies in the production of specialty chemicals, such as pharmaceutical products, several approaches are presented in Bonvin et al. (2006). These approaches deal with operational issues commonly found in industry such as the absence of a steady state and highly non-linear behaviour. Additional challenges include infrequent or delayed on-line measurements of product quality, which is typical for the majority of pharmaceutical operations.

A variety of modelling approaches have been proposed to improve batch operations, these include mechanistic based approaches, such as in Birol et al., 2002 and Goldrick et al. (2015), where the authors used first principle models to found an optimal fed-batch strategy in different penicillin production case studies. Another approach is referred to as Batch–to-Batch (B2B) or run-to-run optimisation. B2B manipulates the conditions from one batch to the next, with the objective of gradually increasing an economic cost function and/or bringing the end-point quality closer to a desired set-point in the presence of disturbances. Recent studies have demonstrated the benefits of using B2B in industrial applications, see for example (Liu et al., 2018), which provides a review of this work. Of particular note to the work proposed in this paper is that of Yabuki et al. (2000), where the authors used mid-correction policies to control the final quality using a predictive model that was developed using a knowledge-based approach. A related study by Camacho et al. (2007) proposed a B2B evolutionary optimization methodology, which they demonstrated was able to significantly increase the end-point quality of a simulated fermentation process, when compared with knowledge based approaches. However, if the behaviour of the disturbances change from batch to batch, then model predictive control (MPC) has been shown to be a more effective technique for ensuring that the end-point quality of the process meets its desired value (Flores-Cerrillo and MacGregor, 2003).

End-point, or run-end MPC uses the available on-line measurements to provide an estimate of the expected end-point quality at regular intervals during the batch. The controller then applies corrective action as and when required, to ensure that the product quality at the end of the batch meets its target. The corrective action applied by the MPC will be the adjustment of the manipulated variable trajectories (MVTs). These trajectories can be manip-

* Corresponding author.
E-mail addresses: carlos.duran@manchester.ac.uk (C.A. Duran-Villalobos), s.goldrick@ucl.ac.uk (S. Goldrick).

https://doi.org/10.1016/j.compchemeng.2019.106620

ulated from the current time, through to the expected end-point of the batch. An extensive review of MPC, specifically focused on its application to chemical engineering applications was reported by Kumar and Ahmad (2012). There have been many MPC strategies that have been proposed for chemical process system, which include the use of classical state-space models to provide quality predictions, such as the article presented by Sheng et al. (2002), where the authors proposed a new generalised predictive controller for systems where sampling is non-uniform. Alternative approaches include the use of Multivariate Statistical Process Control (MSPC) models, such as the one implemented into an end-point controller by Flores-Cerrillo and MacGregor (2003) to regulate particle-size distribution in an emulsion polymerization process.

MSPC refer to a collection of statistical-based techniques that attempt to condense the information contained within large numbers of sensor measurements, into a reduced number of composite variables. In batch control applications, Multi-way Partial Least Squares (MPLS) has been shown to be a powerful regression tool, which in combination with adaptive techniques can be used to provide an approximation of the dynamic characteristics of a batch process with only a limited quantity of data (Joe Qin, 1998). For example, Flores-Cerrillo and Macgregor (2004) demonstrated how MPLS models could be identified for a condensation polymerization process and then used within a cost function, that when solved, using a Quadratic Programming (QP) optimisation approach could adjust the MVTs to improve the consistency of the process. The main advantage of using an MPLS model is that the optimisation could be achieved in the latent variable space of the model, resulting in significantly less computational overheads. A similar approach was employed by Wan et al. (2012) for the control of final quality in a batch process, but their approach also considered hard and soft manipulated variable constraints in the QP problem and applied disturbance rejection control. Their results showed that a disturbance model in the MPLS-based controller improved final quality and that the inclusion of constraints in the manipulated variable in the optimisation problem ensured that the upper and lower bounds were respected. A limitation with the proposed control system was that the soft constraints, used within the QP optimisation formulation, needed to be tuned to ensure that the final quality predictions remained within the score space defined by the identification data-set.

Laurí et al. (2013) demonstrated that in the application of end-point MPC to a fermentation process, there were considerable benefits resulting from the inclusion of hard validity constraints. These constraints were applied to the MPLS model's Hotelling's (T^2) and Square Prediction Error (SPE) during the optimisation of the controller cost function to ensure that the model did not extrapolate too far from the conditions used to identify the model. The same constraints with adaptations to a B2B optimisation strategy were applied in (Duran-Villalobos et al., 2016) for a B2B optimisation, which modified the control strategy presented by Wan et al. (2012) and solved the QP in the real space, whilst including the effect of the projection of the future changes in the MVT to the 'latent' space. The addition of these terms in the control strategy significantly improved its performance; however, the confidence limits used for the constraints (Laurí et al., 2013; Nomikos and Macgregor, 1995a,b Ündey et al., 2003) were not clearly specified and assumed that the data could be approximated by normal and chi-squared distributions for T^2 and SPE respectively, which was only true in specific applications.

The MPLS-based end-point control strategy, proposed in this article, defines confidence limits that are applied to the hard validity constraints used by Duran-Villalobos et al. (2016), which addressing the limitations encountered with the constraints proposed when using in similar control strategies (Laurí et al., 2013; Nomikos and Macgregor, 1995a,b; Ündey et al., 2003). The main differences between the work proposed in this article and similar approaches (Flores-Cerrillo and MacGregor, 2005, 2004; Wan et al., 2012) are that the current approach uses adaptive techniques to improve the model from one batch to the next and that the MVT optimisation is solved not in the score space but in the real space. Appendix A shows a comparison of the results using previous strategies against the proposed approach.

MPLS-based end-point control requires the future progress of the process to be estimated. There are various approaches that have been proposed for achieving this and the accuracy of the resulting controller is very much dependent on the technique chosen. Future estimates of the process variables are determined in the latent variable space and 'missing data' techniques are typically used for doing this. In this article, the capabilities of two such techniques are compared and a novel approach is proposed that integrates two control objectives for the regulation of multiple batch runs. The capabilities of the proposed controller is demonstrated using a benchmark simulation of an industrial penicillin fed-batch fermentation process (Goldrick et al., 2015). Previous studies have demonstrated how fault detection and diagnosis tools can be applied to this simulated process (Luo and Bao, 2018). However, there have been no studies that have applied model-based control techniques to it.

The two objectives of the controller proposed in this article are to: 1. Reach an optimal final penicillin concentration in a B2B optimisation campaign, beginning with an a-priori trajectory for the primary manipulated variable (glucose feed); 2. Reduce variability in the final penicillin concentration by adjusting the glucose feed trajectory within the batch using MPC.

The structure of this paper begins with an overview of the industrial penicillin simulation and operation methodology in Section 2. MPLS and its identification and adaption from one batch to the next is defined in Section 3. The two control objectives are then formulated in Section 4 and the cost function and QP solution is described in Section 5. The results of the B2B optimisation and end-point MPC control when applied to the simulation is presented and discussed in Section 6. Finally, conclusions are provided in Section 7.

2. Case study

Regarding the test and comparison of alternative strategies for industrial control, Bonvin (1998) holds the view that there is a definite need for realistic benchmarks and that the developed control strategies should not be oversold but rather evaluated experimentally on pilot-plant and industrial reactors. A notable example of a realistic simulation of an industrial fermentation process is presented in Goldrick et al. (2015). This simulation (IndPenSim: www.industrialpenicillinsimulation.com), available in MATLAB, describes a complex mechanistic model of a penicillin fermentation process that has been validated using data collected from an industrial process. The industrial process was a 100,000 l bioreactor, which produced the *Penicillium chrysogenum* strain.

The main simulation parameters that were used in both the B2B and MPC campaigns described in this article are provided in Table 1.

IndPenSim includes random variations in the initial conditions for several variables, including initial volume and seed concentrations. The simulation also includes within-batch variation in the penicillin specific production rate, biomass specific growth rate, substrate concentration, acid/base concentration, phenylacetic acid concentration, coolant inlet temperature and oxygen inlet concentration. The addition of disturbances in the simulation seeks to present a more realistic challenge, with similar variability in process parameters typically encountered in industrial operation.

Table 1
IndPenSim simulation parameters for the B2B and MPC campaigns.

Simulation parameter	B2B	MPC
Batch total time	230 h	230 h
Control action interval	230 h	10 h
Start of the control action	1 h	50 h
Optimal Penicillin conc.	30 g/L	30 g/L
Campaign length	50 batches	80 batches
Measurements interval	1 h	1 h

Table 2
IndPenSim simulation parameters used in the MPLS model.

Input variable	Initial condition	Initial variability (+/−)
CO_2 conc. Off gas	0.038%	0.001%
DO_2 conc.	15 mg/l	0.5 mg/l
O_2 conc. Off gas	0.02%	0.05%
Penicillin conc.	0 g/l	0 g/l
pH	6.5 (−)	0.1 (−)
Temperature	297 (K)	0.5 (K)
Volume	5.8e4 l	500 l

Table 2 shows the nominal values for the process parameters that were used to identify the MPLS model in this work.

The simulations and control strategies for the B2B and MPC campaigns were implemented in Matlab R2017a, utilizing the *Global Optimization* and *Optimization* toolboxes.

3. MPLS model identification

The control strategies presented in this article use an MPLS model that is extensively described in Duran-Villalobos et al. (2016). However, this section will present a brief description of the model identification process for clarity.

3.1. PLS regression

PLS regression is a multivariate statistical technique where a linear regression model is found by projecting the predictor, $\boldsymbol{X}$, and response, $\boldsymbol{Y}$, variables into orthonormal vectors in a 'Latent Variable' (LV) space, which explains the maximum covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$. In contrast with standard regression techniques, this regression is particularly well suited when the matrix of predictors, $\boldsymbol{X}$, presents high multicollinearity among its values, such as measurements over time of typical fermentation processes.

Eqs. (1) and (2) show the bi-diagonal PLS model proposed my Martens and Naes (1989).

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E} \tag{1}$$

$$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Q}^{\boldsymbol{T}} + \boldsymbol{F} \tag{2}$$

where the matrix of scores, $\boldsymbol{T}$, contains the values of each row (observations) of $\boldsymbol{X}$ in the LV space. The matrix of loadings, $\boldsymbol{P}$ and $\boldsymbol{Q}$, contain the projection of each column of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, in the LV space. And the matrix of residuals, $\boldsymbol{E}$ and $\boldsymbol{F}$, are matrices of residuals between the regression and the data in the identification set.

The matrix of responses, $\boldsymbol{Y}$, can be defined as a vector, $\boldsymbol{y}$, for end-point qualities, such as final penicillin concentration. The work presented in this article assumes that measurements of the response variable are only available at the end of the batch. As a result, the estimated value of a response for a new batch, i, can be described as in Eq. (3).

$$\hat{y}_i = \boldsymbol{t_i}\boldsymbol{Q}^{\boldsymbol{T}} \tag{3}$$

where the score vector, $\boldsymbol{t_i}$, for a new batch can be obtained by projecting the new vector of measurements, $\boldsymbol{x_i}$, into the projection weight matrix, $\boldsymbol{W}$, as shown in Eq. (4)

$$\boldsymbol{t_i} = \boldsymbol{x_i}\boldsymbol{W} \tag{4}$$

3.2. Data structure

The measured variables in the identification data set, which contains measured variables (of size J), time intervals (of size K) and batch number (of size I); are transformed into a 2-dimensional array as shown in Eq. (5). This transformation allows the PLS model to capture time varying dynamics within multivariate data (Nomikos and MacGregor, 1995a).

$$\boldsymbol{X_{3D}} \in \mathbb{R}^{I\ \times J\times K} \rightarrow \boldsymbol{X_{2D}} \in \mathbb{R}^{I\ \times JK} \tag{5}$$

In addition, the vector of measurements at each new batch, $\boldsymbol{x_i}$, the matrix of weights, $\boldsymbol{W}$, and the matrix of loadings, $\boldsymbol{P}$, are divided as shown in Eqs. (6)–(8).

$$\boldsymbol{x_i} = \left[\boldsymbol{x_p}\ \boldsymbol{u_n} + \Delta\boldsymbol{u}\ \boldsymbol{x_f}\right] = \left[\boldsymbol{x_{pu}}\ \boldsymbol{x_f}\right] \tag{6}$$

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W_p} \\ \boldsymbol{W_u} \\ \boldsymbol{W_f} \end{bmatrix} = \begin{bmatrix} \boldsymbol{W_{pu}} \\ \boldsymbol{W_f} \end{bmatrix} \tag{7}$$

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P_p} \\ \boldsymbol{P_u} \\ \boldsymbol{P_f} \end{bmatrix} = \begin{bmatrix} \boldsymbol{P_{pu}} \\ \boldsymbol{P_f} \end{bmatrix} \tag{8}$$

where $\boldsymbol{u_n}$ is a vector containing the nominal values for the MVT, Δu is a vector containing the optimal change in the MVT. The subscripts represent: $\boldsymbol{p}$ past horizon for measurements, $\boldsymbol{u}$ control horizon for the MVT, $\boldsymbol{p_u}$ past and control horizon for the MVT, and $\boldsymbol{f}$ the prediction/future horizon for the measurements.

3.3. Model adaptation

The MPLS model used in this work is updated at the end of each batch that the control system is applied. The objective of this update, which is achieved using the recursive techniques proposed in Dayal and MacGregor (1997) and Joe Qin (1998), is to allow the controller to track the dynamics of the process as the operating conditions vary as a consequence of the changes imposed by the B2B optimiser and to 'refine' the MPLS model used within the MPC, This adaptation is necessary because the PLS regression can represent only the linear dynamics of the process local to the region of operation that has been used to identify the model. The recursive technique employed in this article is shown in Eq. (9).

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{\lambda X_{i-1}} \\ \boldsymbol{x_i} \end{bmatrix} \textit{ and } \boldsymbol{y} = \begin{bmatrix} \boldsymbol{\lambda y_{i-1}} \\ \mathrm{y}_i \end{bmatrix} \tag{9}$$

where a forgetting factor, λ, is applied to the data collected from previous batches to ensure that the model forgets the behaviour of historical batches, but remembers the most recent batches. This allows the controller to follow substantial changes in the process dynamics, which may be non-linear. The approach for selecting a suitable value for λ is formulated in Duran-Villalobos et al. (2016) and it must be chosen such that the number of batches is relevant to the conditions around which the process is currently operating. In the work presented in this article, the dynamics of the process did not seem to change substantially through the B2B and MPC campaigns, as changes to the value of λ did not offer any significant improvement to the prediction accuracy of the MPLS model. Therefore, λ was set to a value of 1 for all the studies presented in this article.

3.4. Number of latent variables

The dimension of the LV space is usually determined by cross-validation (CV) to ensure that the resulting model provides a robust prediction of the response variables (Camacho et al., 2007). A commonly applied method for determining the dimension of the LV space for a PLS models is leave-one-out cross-validation (Martens and Naes, 1989). However, this method can result in unnecessarily large models that can introduce increased risks of overfitting. This was demonstrated by Xu and Liang (2001), where the results obtained from a multivariate simulation study showed that a method known as Monte Carlo Cross-Validation (MCCV) provided improved performance when compared with leave-one-out cross-validation.

As a consequence of these results, MCCV was used to find the number of LVs for each of the models used in this work. The MCCV approach determines the appropriate number of LVs, A, by randomly drawing a collection of observations, of size v, and using these observations to identify a PLS model. This process is repeated N times for each number of LVs, a, as shown in Eq. (10). Then, the results for each value of a are compared and the one with the minimum value is selected to be A. A basis to select the value of v and N are presented in Xu and Liang (2001).

$$MCCV(a) = \frac{1}{Nv}\sum_{i=n}^{N} \|\boldsymbol{y}_{v,\ n} - \widehat{\boldsymbol{y}_{v,n}}\| \tag{10}$$

4. Control objectives

4.1. B2B optimisation

The first control objective that was applied in this work was initially formulated in the article by Duran-Villalobos et al. (2016). This technique attempts to bring the end-point quality (final penicillin concentration for the case study considered in this work) closer to the desired set-point by allowing the B2B optimiser to make adjustments to the MVT (which in this work is the glucose feed rate). These MVT adjustments were made through consideration of the data collected from previous batches, which were used to improve the accuracy of an adaptive MPLS model.

Fig. 1 shows a simplified flowchart of the iterative control strategy used within the B2B optimiser. First, the plant is excited with a Pseudo Random Binary Signal (PRBS) that was passed through a low-pass filter and then added to a pre-optimised MVT over a small number of batches. Duran-Villalobos et al. (2016) showed that 3–8 batches provided enough data to obtain an MPLS model with sufficient accuracy and similar results were obtained in this work, although it is expected that this is likely to be problem dependent. The MPLS model is then used within the QP cost function that is solved to find an optimal MVT that minimises the difference between the desired and predicted end-point qualities. The optimised MVT is filtered, using a low-pass Finite Impulse Response filter with a cut-off frequency of 10% of the maximum frequency as this has been shown to be beneficial in previous studies (Camacho et al., 2015, 2007; Duran-Villalobos et al., 2016). This filtered MVT is excited with low-amplitude PRBS (3%) and used throughout the subsequent batch, and finally, the data collected from this new batch is used to update the model and the next batch run.[1]

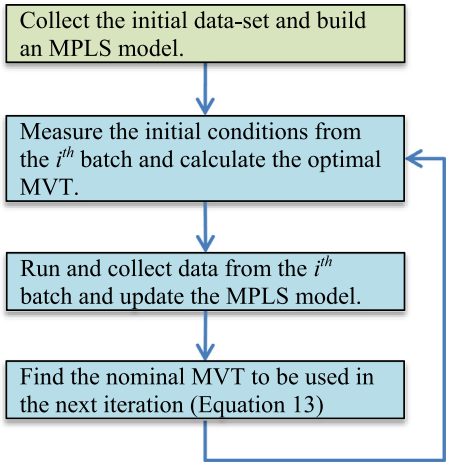

Fig. 1. Batch to batch (B2B) optimisation flowchart.

A major difference to the approach applied in this work, compared with the techniques proposed by Duran-Villalobos et al. (2016) is that the proposed approach, determines whether or not the current optimised MVT will bring the process closer to the set-point for the end-point product quality. This addition was found to improve significantly the convergence speed of the B2B optimiser. Eqs. (11)–(13) show how the decision is carried out from batch to batch.

$$e_{past} = min\left((\boldsymbol{y} - y_{sp})^2 \right) \tag{11}$$

$$e_i = (y_i - y_{sp})^2 \tag{12}$$

$$\boldsymbol{u}_{i+1} = \begin{cases} \mathbf{u_i} \text{ if } e_i < e_{past} \\ \mathbf{u_{epast}} \text{ if } e_i > e_{past} \end{cases} \tag{13}$$

where e_{past} is the minimum quadratic error found between the measured end-point qualities, $\boldsymbol{y}$, and the desired set-point y_{sp}; e_i is the quadratic error between the ith end-point quality, y_i, and the set-point, y_{sp}; $\boldsymbol{u_i}$ is the MVT for the ith batch and $\boldsymbol{u_{epast}}$ is the MVT corresponding to e_{past}.

One of the main challenges in an industrial control strategy is to deliver the set-point in the presence of disturbances and potentially to changes in the dynamics of the process. If any of the disturbances are highly correlated from one batch to the next, the information from previous batches can be used in the B2B optimisation to determine how the current batch should be operated to mitigate any similar disturbances. However, if the behaviour of the disturbances is stochastic and changes from batch to batch, then within-batch control, using techniques such as MPC, is recommended (Flores-Cerrillo and MacGregor, 2003).

4.2. MPC

To ensure that the set-point is met and that variation in product quality is reduced, MPC is applied. The MPC strategy adjusts the MVT at different control action points through the batch, solving the same QP problem as the B2B optimiser and using a similar data-driven adaptive MPLS model.

Fig. 2 shows a simplified flowchart of the iterative control strategy for the MPC. First, the iterative process uses the same strategy as the B2B optimiser: the plant is excited with a low filtered PRBS and an MPLS model is identified; however, when MPC is applied, a PRBS is applied to a 'golden' trajectory which has been previously optimised. Then, the MPC strategy is executed using a nested loop structure.

[1] The reader is invited to read Duran-Villalobos et al. (2016) for further details.

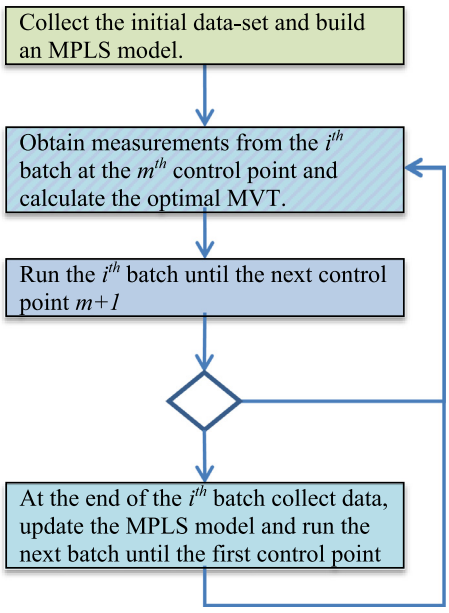

Fig. 2. Model Predictive Control (MPC) flowchart.

The inside control loop calculates the future optimal changes to the MVT, $\Delta \boldsymbol{u}$, at each mth control point until the batch is complete. This calculation is similar to the strategy presented in Flores-Cerrillo and MacGregor (2005), where a QP searches for the score value that keeps the change in the score closer to those used in the identification dataset. The MVT is then obtained by projecting this new optimised score into a Multivariate Principal Component Analysis (MPCA) model. In contrast, the MPC strategy described in this article calculates directly the necessary changes that should be applied to the MVT within the QP. This QP, described in Section 5, includes a cost function which has the objective of minimizing the difference between the future predicted end-point quality and its set-point. In addition, the constraints imposed on the QP ensure that the changes in the MVT remain within the physical capabilities of the plant and the score space region used to identify the MPLS model.

Once the optimal steps for the MVT are calculated, at control point, the batch is run until the next control point, $m+1$, were the data vector $\boldsymbol{x_i}$, used to predict the end-point quality, is updated with the new process measurements to calculate the future optimal steps at the $m+1$ control point. This process is repeated at each control point until the end of the batch.

At the end of each batch, the outer control loop is implemented: the data from the batch, i, is collected and the MPLS model is updated. Then batch, $i+1$, is run until the first control point using the same 'golden' trajectory for the MVT. At this stage, the control loop is executed again to obtain the optimal MVT at each control point until the end of batch, $i+1$. This process is repeated for each batch until the end of the MPC campaign.

During the MPC experiments, it was found that at each control point, the available measurements improved the quality of the model's predictions during every batch. It was also found that the predictive accuracy of the model was reduced when longer intervals in the control action were used, which is expected (Bonvin et al., 2006).

In the results presented in this article, the control interval was set to 10 h. Reducing the control interval below this was not found to improve the performance of the control system. The most suitable value for the control interval will be dependent upon the dynamics of the process being studied.

5. QP optimisation

5.1. Cost function

As explained in Section 4, the optimisation of the MVP seeks to bring the end-point quality closer to the set-point, while respecting the limits in the decision space defined by the data sued to identify the MPLS model. This objective was formulated in Duran-Villalobos et al. (2016), where a widely used cost function (Qin and Badgwell, 2003) to be minimised is defined as a trade-off between the square of the error between the set-point and the predicted output and the square of any changes made to the MVT. This cost function is presented in Eq. (14).

$$\min_{\Delta \boldsymbol{u}} \left(\hat{y}_i - y_{sp}\right)^T \left(\hat{y}_i - y_{sp}\right) + \Delta \boldsymbol{u}^T \boldsymbol{M} \Delta \boldsymbol{u}$$
$$s.t. \begin{cases} \hat{y}_i = \boldsymbol{t_i} \boldsymbol{Q}^T \\ lb \leq \boldsymbol{u_n} + \Delta u \leq ub \\ Vol_{ini} + \Delta_{Vol} \leq Vol_{max} \\ J_e \leq 1 \ and \ \ J_t \leq 1 \end{cases} \tag{14}$$

where:

- The cost function includes a diagonal matrix of weights, $\boldsymbol{M}$, that is used to moderate the change in the MVT made by the QP optimisation. The diagonal elements in the matrix of weights, $\boldsymbol{M}$, were set to a value of 0.01. This value was found to be a good trade-off between the convergence speed and aggressiveness of the control action in the MPC. However, in the B2B optimisation any value less than 0.1 was found to reach the same convergence speed.
- The cost function is subject to the calculation of the estimated end-point quality expressed in the first constraint. However, as the values of the future process measurements are unknown, when the control action is implemented, the score value for the current batch needs to be estimated using the missing data techniques, described in Section 5.2. The second constraint introduces limits to the MVT, which are imposed through physical restrictions to the magnitude of the feed-rate. These constraints, mean-centred, are represented by the lower and upper bound vectors, $\boldsymbol{lb}$ and $\boldsymbol{ub}$.
- The third constraint ensures that the feed rate restricts the maximum volume in the reactor to below the physical limit imposed by the vessel, which in this work was 100,000 litres. This constraint is briefly described in Section 5.3.
- Finally, the validity constraint limits J_e and J_t restrict the solution space over which the QP optimisation searches to ensure that the solution is within the space of the data used to identify the MPLS model. Validity constraints were proposed in Duran-Villalobos et al. (2016) and Laurí et al. (2013). However, in this article, the confidence limits that are applied do not assume that the data follows a normal distribution, which is explained further in Section 5.4.

In the work presented in Duran-Villalobos et al. (2016), the predicted values of the future measurements were used to calculate the effect that the change in the MVT had on the predicted output. In contrast, in this work, the score for the optimised batch is defined as the sum of the estimated score vector for the past measurements $\widehat{\boldsymbol{t_i}}$, and the effect in the score change in the MVT, as in Eq. (15). Using this strategy, the value of future measurements is unnecessary for the calculation of the predicted output.

$$\boldsymbol{t_i} = \widehat{\boldsymbol{t_i}} + \Delta \boldsymbol{u} \boldsymbol{W_u} \tag{15}$$

By substituting Eq. (15) into Eq. (14), the QP problem, to optimise the MVT, can be expressed as shown in Eq. (16).

$$\min_{\Delta \boldsymbol{u}} \ \tfrac{1}{2}\Delta \boldsymbol{u}^T \boldsymbol{H} \Delta \boldsymbol{u} + \boldsymbol{f}^T \Delta \boldsymbol{u} \quad s.t. \begin{cases} \boldsymbol{H} = \boldsymbol{W_u Q^T Q W_u^T} + \boldsymbol{M} \\ \boldsymbol{f}^T = \left(\widehat{\boldsymbol{t}}_{\boldsymbol{i}} \boldsymbol{Q}^T - \boldsymbol{y_{sp}}\right) \boldsymbol{Q W_u^T} \\ \boldsymbol{lb - u_n} \le \Delta \boldsymbol{u} \le \boldsymbol{ub - u_n} \\ Vol_{ini} + \Delta_{Vol} \le Vol_{max} \\ J_e \le 1 \ and \ J_t \le 1 \end{cases} \tag{16}$$

5.2. Estimation with missing data

To optimise the MVT, it is necessary to estimate the future end-point quality. However, the values of the future measurements necessary to obtain the score vector, $\boldsymbol{t_i}$, are not available at each control point. To solve this problem, Flores-Cerrillo and MacGregor (2004) proposed the use of missing data algorithms to estimate the scores using an existing MPCA or MPLS model. These predictions are possible because such models capture the time varying structure of the data over the entire batch trajectory, described by the covariance of the measurements vector.

There have been many missing data estimation techniques that have been proposed (Arteaga and Ferrer, 2002), with the Projection to the Modal Plane (PMP) and the Trimmed Score Regression (TSR) often cited as the most suitable methods (Arteaga and Ferrer, 2002; Ed, 2013; García-Muñoz et al., 2004; Gins et al., 2009; Vanlaer et al., 2011). .

In the PMP method, the score is estimated by regressing the vector of known measurements and nominal values of the MVT into the score space defined by their respective loading matrix values, $\boldsymbol{P_{pu}}$, in the identification dataset. The equation to obtain a prediction of the score value, using PMP, is shown in Eq. (17).

$$\widehat{\boldsymbol{t}}_{\boldsymbol{i}} = \begin{bmatrix} \boldsymbol{x_p} & \boldsymbol{u_n} \end{bmatrix} \boldsymbol{P_{pu}} \left(\boldsymbol{P_{pu}^T P_{pu}}\right)^{-1} \tag{17}$$

If the MPLS model is identified using the diagonal method proposed by Wold et al. (1987) instead of the bi-diagonal method proposed by Martens (2001), then the matrix of weights, $\boldsymbol{W}$, must be used instead of the loading matrix, $\boldsymbol{P}$, (Nelson and Taylor, 1996).

In contrast, the TSR method seeks to reconstruct the score $\boldsymbol{T}$ from the trimmed scores $\boldsymbol{T_{pu}}$ through a least squares estimator matrix $\boldsymbol{B}$, as shown in Eq. (18).

$$\boldsymbol{T} = \boldsymbol{T_{pu}} \boldsymbol{B} = \boldsymbol{X_{pu}} \boldsymbol{W_{pu}} \boldsymbol{B} \tag{18}$$

This regression model, can then be used to estimate the score vector for a new batch with missing measurements as shown in Eq. (19).

$$\widehat{\boldsymbol{t}}_{\boldsymbol{i}} = \begin{bmatrix} \boldsymbol{x_p} & \boldsymbol{u_n} \end{bmatrix} \boldsymbol{W_{pu}} \left(\boldsymbol{W_{pu}^T X_{pu}^T X_{pu} W_{pu}}\right)^{-1} \times \boldsymbol{W_{pu}^T X_{pu}^T X W} \tag{19}$$

This method is equivalent to the PMP method if the data matrix, $\boldsymbol{X}$ is of rank $\boldsymbol{A}$ and, after extracting all the LV there is no error remaining (Arteaga and Ferrer, 2002).

5.3. Volume constraints

The constraint for keeping the volume below the maximum capacity is similar to the one proposed in Duran-Villalobos et al. (2016). However, the case study simulation contains multiple feeds and discharge rates in addition to an evaporation rate, which all affect the volume. Consequently, the impact of the other variables affecting volume, $E\,(\Delta V)$, were also considered, as shown in Eq. (20).

$$\sum_{k=1}^{K} \boldsymbol{u_n}(k) + \Delta \boldsymbol{u}(k) \le V_{max} - V_{ini} - E(\Delta V) \tag{20}$$

where the expected value of other variables impacting the volume $E(\Delta V)$ are defined as the average value of the final volume in the initial identification dataset. The reason for this, is that in a real application, the precise values of some of the variables affecting volume, such as evaporation rate, may not be known.

Following the mean-centring of the data prior to the MPLS model identification, Eq. (20) can be written as Eq. (21).

$$\boldsymbol{i_u S_u} \Delta \boldsymbol{u} \le V_{max} - V_{ini} - \mu_V - \boldsymbol{i_u}(\boldsymbol{\mu_u} + \boldsymbol{S_u u_n}) \tag{21}$$

where $\boldsymbol{i_u}$ is a vector of ones with the same length as the MVT, $\boldsymbol{S_u}$ is a diagonal matrix of the MVTs standard deviation in the identification dataset and μ_V is the average value of the final volume in the identification dataset.

Having defined E (ΔV) as the mean value in the identification dataset, μ_V, introduces a small error into the calculation. However, despite this, the proposed methodology obtained good approximations of volume in the case study investigated in this article, ensuring that the constraints were respected. This can be illustrated in Fig. 3, which shows the final volume in the vessel for 50 batches that were part of a B2B campaign.

5.4. Validity constraints

Validity constraints were included to restrict the score space of the QP solution into a region described by data collected from the batches used to identify the MPLS model. A useful methodology for this purpose are the hard validity constraints presented in Laurí et al. (2014), where constraints are imposed on the Hotelling's (T^2) and the Q statistics.

The T^2-statistic based validity indicator, J_t, is shown in Eq. (22). This validity indicator measures the deviation of $\boldsymbol{t_i}$ from the score region covered by the identification dataset.

$$J_t = \frac{\boldsymbol{t_i} \left(\boldsymbol{S_\alpha^2}\right)^{-1} \boldsymbol{t_i^T}}{J_{tmax}} \tag{22}$$

where $(\boldsymbol{S_\alpha^2})^{-1}$ is a diagonal matrix that contains the covariance of each LV in the score matrix, $\boldsymbol{T}$; and J_{tmax} provides a normalization variable for J_t in the identification dataset.

The Q-statistic based validity indicator, J_e, is shown in Eq. (23). This validity indicator provides a measure of the error between the predictor vector, $\boldsymbol{x_i}$, and its reconstructed value from the MPLS model.

$$J_e = \frac{\boldsymbol{e_i e_i^T}}{J_{emax}} \tag{23}$$

where $\boldsymbol{e_i}$ is the squared error of projection of the predictor variables; and J_{emax} provides a normalization variable for J_e in the identification dataset.

The squared error of projection for the QP optimisation, formulated in Duran-Villalobos et al. (2016) can be reformulated as shown in Eq. (24) by adding the effect of $\Delta \boldsymbol{u}$ on the future measurements.

$$\boldsymbol{e_i} = \widehat{\boldsymbol{x}}_{\boldsymbol{i}} \left(\boldsymbol{I - W P^T}\right) + \Delta \boldsymbol{u} \left(\boldsymbol{I - W_u P_u^T}\right) + \Delta \boldsymbol{u} \boldsymbol{\theta} \left(\boldsymbol{I - W_f P_f^T}\right) \tag{24}$$

where the future measurements can be estimated by projecting the estimated score vector, obtained from the missing data algorithms, into the loadings values corresponding to the future measurements. As a result, the vector of predictors, $\widehat{\boldsymbol{x}}_{\boldsymbol{i}}$, can be constructed as shown in Eq. (25).

$$\widehat{\boldsymbol{x}}_{\boldsymbol{i}} = \begin{bmatrix} \widehat{\boldsymbol{x}}_{\boldsymbol{i}} & \boldsymbol{u_n} & \widehat{\boldsymbol{t}}_{\boldsymbol{i}} \boldsymbol{P_f^T} \end{bmatrix} \tag{25}$$

The estimator for the effect of $\Delta \boldsymbol{u}$ on the future measurements, $\boldsymbol{\theta}$, is obtained from the PMP and TSR missing data algorithms as shown in Eq. (26).

$$\boldsymbol{\theta_{PMP}} = \boldsymbol{P_u} \left(\boldsymbol{P_u^T P_u}\right)^{-1} \boldsymbol{P_f^T} \quad Or \quad \boldsymbol{\theta_{TSR}} = \boldsymbol{W_u} \left(\boldsymbol{W_u^T X_u^T X_u W_u}\right)^{-1} \boldsymbol{W_u^T X_u^T X W P_f^T} \tag{26}$$

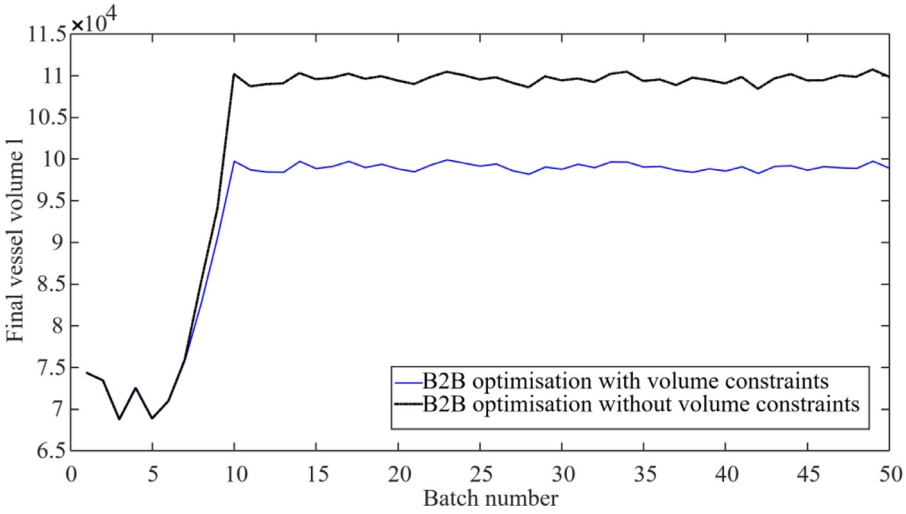

Fig. 3. Volume constraints for a B2B optimisation campaign.

As previously stated, J_{tmax} provides a normalization parameter for the T^2-statistic based validity indicator. This normalization can be represented as a confidence interval from which we want to keep the decision space in the QP optimisation. A well-known upper confidence limit for the T^2-statistic was described in Qin (2003), where the author presented an upper control limit that can be well approximated as a chi-squared distribution. This approximation is only possible under the condition that the data follows a multivariate normal distribution. Then the normalization variable, J_{tmax}, can be defined by Eq. (27).

$$J_{tmax} = \chi^2_{A,\alpha} \tag{27}$$

where χ^2 is a chi-squared distribution with A degrees of freedom and for a significance level α (where the tolerance is $1-\alpha$).

On the other hand, J_{emax} provides a normalization parameter for the Q-statistic based validity indicator. This normalization can also be represented as a confidence interval from which we want to keep the decision space in the QP optimisation. A good example for an upper control limit for Q was formulated in Jackson et al. (1979). This control limit, under the assumption that the data follows a normal distribution, was calculated from a Gaussian approximation of a normal distribution. Therefore, the normalization variable, J_{emax}, can be defined by Eq. (28).

$$J_{emax} = \delta^2_{A,\alpha.} \tag{28}$$

where χ^2 is the approximation of a normal distribution, defined in Qin (2003).

The assumption that the data follows a normal distribution is a drawback of these control limits since not all process data will present this characteristic. For instance, in the B2B optimisation campaign presented in this article, the probability distribution of the data is constantly changing since the end-point quality is changing from one batch to the next. This change in the observations is not random, and causes the data to have a non-normal distribution. In this work, this was confirmed using a Kolmogorov-Smirnov test (Ramani, 1974).

A more general approach to defining confidence intervals, which do not make assumptions as the distribution of the data is described in Desharnais et al. (2015), where the author uses a bootstrap-resampling technique to calculate confidence intervals in non-normal datasets. This technique infers confidence intervals from an empirical distribution function, assuming that the collected data, having being drawn from the population, are the best available representatives of the population. The confidence intervals are estimated from 'resampled' datasets, which are formed by individual randomly chosen samples from the original dataset.

The normalized variables J_{tmax} and J_{emax} can then be defined, respectively, as shown in Eqs. (29) and (30)

$$J_{tmax} = \beta_{nb,\alpha} diag\left(\boldsymbol{T}\left(\boldsymbol{S}^2_{\boldsymbol{\alpha}}\right)^{-1}\boldsymbol{T}^{\boldsymbol{T}}\right) \tag{29}$$

$$J_{emax} = \beta_{nb,\alpha} diag\left(\boldsymbol{E}\boldsymbol{E}^{\boldsymbol{T}}\right) \tag{30}$$

where β is the upper confidence interval calculation using bootstrap-resampling (Desharnais et al., 2015) and nb is the number of resample datasets (typically 1000–10,000).

By using Eqs. (27)–(30), the validity constraints can then be formulated as the nonlinear inequality constraints shown in Eqs. (31) and (32).

$$\frac{J_t}{J_{tmax}} \leq 1 \tag{31}$$

$$\frac{J_e}{J_{emax}} \leq 1 \tag{32}$$

Despite the fact that both Q and T^2 statistics are used for process monitoring, it is necessary to point out that they provide different roles in process monitoring. The Q-statistic measures the predictors' correlation consistency of a certain batch with the identification data-set, while the T^2-statistic measures the distance to the origin in the LV subspace.

6. Results and discussion

The results shown in this section use the same starting seed for the random number generator that was used in Matlab to introduce variability into the process. This allows an accurate comparison to be made of different approaches for a given control campaign, since the generated random numbers are the same for each approach.

The initial pre-optimised feed, used to identify the initial MPLS model consisted of 5 batches with a nominal feed trajectory for both B2B and MPC campaigns. This nominal feed trajectory consisted of a gradual increase from 0 l/h to 50 l/h for the first 4 h and then a constant value of 50 l/h for the remainder of the batch. A filtered (Duran-Villalobos et al., 2016) PRBS of +/−25 l/h was then added to the constant feed.

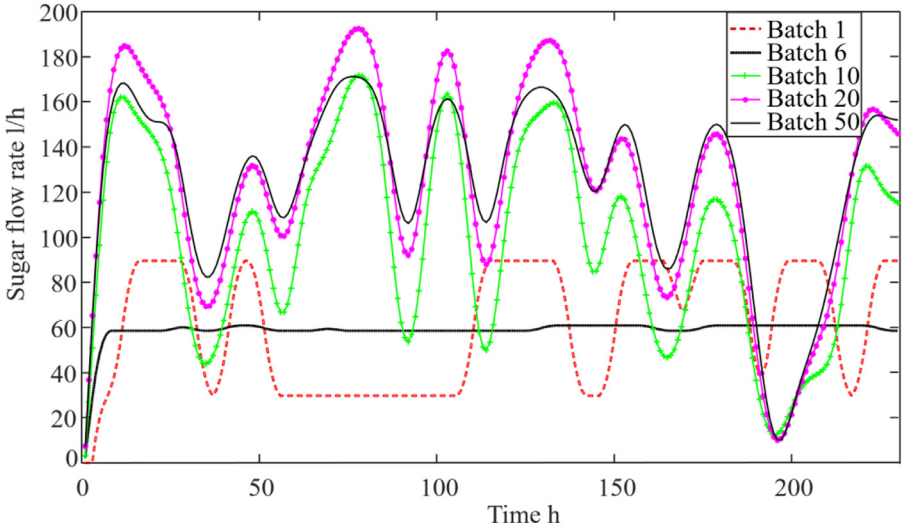

Fig. 4. MVT progression for a B2B-TSR optimisation campaign.

Other parameters used in the experiments were:

- Low-pass filter characteristics: Zero–phase low pass Finite Impulse Response (FIR) filter with a cut-off frequency of 10% of the maximum frequency (Nyquist frequency).
- Lower bound constraints for the actuator = =0 l/h.
- Upper bound constraints = =200 l/h.
- *nb* for the bootstrap calculation = =2000.
- Confidence tolerance of 97%, therefore α = =0.03.

6.1. Validity constraints in the B2B campaign

The objective of the B2B optimisation was to bring the final penicillin concentration (end-point quality) to a pre-established set-point (30 g/l), by optimising the trajectory of the sugar feed flow rate (manipulated variable) from one batch to the next. An important aspect of the work presented in this article was to observe the effect of the validity constraints in the B2B campaign.

Fig. 4 shows how the trajectory of the substrate feed changed during a typical B2B campaign. What stands out in this figure is the sharp change in the amplitude of the MVT at the beginning of the optimisation (from batches 1 to 6, and from batches 6 to 10) and a very moderate change afterwards (from batches 20 to 50). In other words, the trajectory converges relatively quickly for the first 10 batches.

The results for the final penicillin concentration of the B2B campaign were collected from 30 replicated experiments to observe the effect of random variability in the progression of the yield improvement. Each replicate, with different initial random seeds, collected the results from 50 batches.

Fig. 5 shows the average final penicillin concentration of 30 replicates for a nominal run and 3 different validity constraint arrangements, of the B2B optimisation campaign using the TSR missing data technique. In this figure we can observe that the fastest convergence and highest yield is achieved when only the validity indicator for the Q-statistic, J_e is applied. Fig. 5 also shows that with this validity constraint the final penicillin concentration increased gradually to approximately 30 g/l.

Table 3 shows several important evaluation parameters taken from the results displayed in Fig. 5 and compares them with the results if open-loop control was applied. The second column shows the mean of the final penicillin concentration averaged over each of the 30 replicates for the 50th batch. The third column shows the final penicillin concentration standard deviation of the 30 replicates for the 50th batch. Finally, the fourth column shows the batch at which the final penicillin concentration converges a maximum regular value.

From the results in Fig. 5 and Table 3 we can observe that when using only J_e, the process converges to a value closer to the set-point (30 g/l) than when other constraints were used or when the process is operated in open-loop. This configuration also has the lowest standard deviation at the end of the campaign. By contrast, the configuration including both validity constraints has the lowest convergence speed and the highest standard deviation.

This high variability caused by the validity constraint imposed on the T^2-statistic, J_t, can be explained by the wide non-stationary range of the identification data that is used at each MPLS model update. This was also observed by Qin (2012), who states that the limits on T^2 are not reliable in practice when the scores from process data do not follow the assumptions of multivariate normality. Therefore, limits on Q may reduce type I and type II errors compared with limits on T^2 (Qin, 2003). This was corroborated with the experimental results of the MPC, which did not show this detrimental effect.

Regarding the use of confidence intervals when assuming a normal distribution in the dataset, Fig. 6 shows the average final penicillin concentration of 30 replicates, when the 3 different confidence intervals were applied to the B2B optimisation campaign

Table 3
Evaluation parameters for the B2B-TSR campaign under different validity constraints configurations.

Control methodology	Yield: mean of batch 50	Dispersion: standard deviation of batch 50	Batch at which convergence achieved
Nominal run (Open-loop)	21.84 g/l	1.22 g/l	n/a
B2B-TSR without validity constraints	29.31 g/l	1.83 g/l	≈Batch 20
B2B-TSR $J_e < 1$	30.12 g/l	1.63 g/l	≈Batch 15
B2B-TSR $J_e < 1$ $J_t < 1$	27.16 g/l	4.98 g/l	≈Batch 30

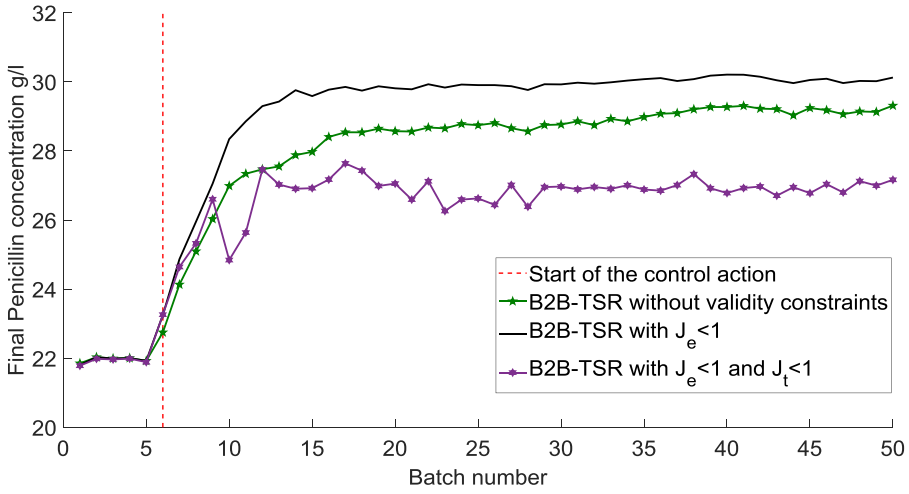

Fig. 5. Final penicillin concentration mean for the B2B-TSR campaign under different validity constraints configurations.

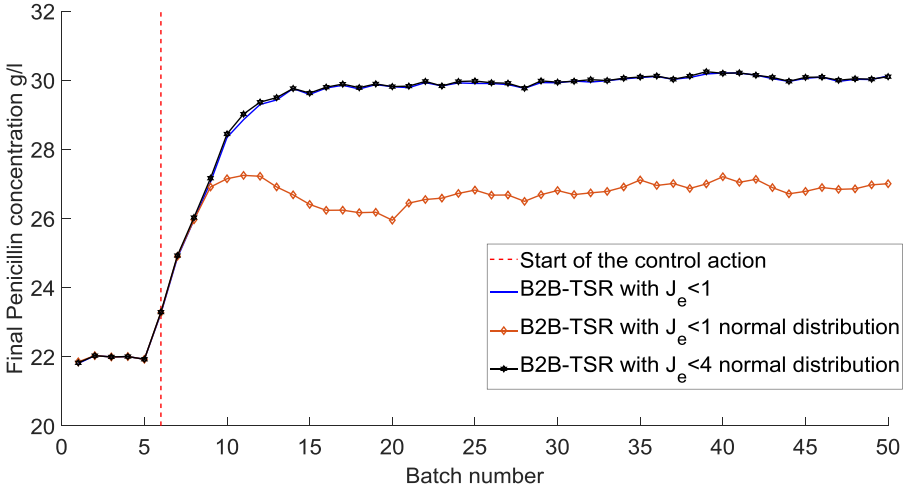

Fig. 6. Final penicillin concentration mean for the B2B-TSR campaign under different confidence intervals methodologies.

Table 4
Evaluation parameters for the B2B-TSR campaign under different confidence intervals methodologies.

Control methodology	Yield: mean of batch 50	Dispersion: standard deviation of batch 50	Batch at which convergence achieved
B2B-TSRJ_e < 1	30.12 g/l	1.63 g/l	≈Batch 15
B2B-TSR Normal dist.J_e < 1	27.00 g/l	4.80 g/l	≈Batch 35
B2B-TSR Normal dist.Je < 4	30.10 g/l	1.64 g/l	≈Batch 15

using TSR. The methodologies that assumed a normal distribution used Eqs. (27) and (28) for the confidence intervals, while the methodologies that did not assume a normal distribution used Eqs. (29) and (30).

Table 4 shows the same evaluation parameters as Table 3 taken from the results displayed in Fig. 6, the B2B-TSR campaign under different confidence interval methodologies.

From Fig. 6 and Table 3 we can observe that the results from the configuration which assumed the dataset to have a normal distribution and have the validity constraint $J_e < 1$ has a much lower convergence speed and higher standard deviation than the results from the configuration which did not assume the dataset to have a normal distribution. The results of the latter, are similar to those which have the validity constraint $J_e < 4$ and assumed the dataset to have a normal distribution. This result suggests that the QP optimisation space is too constrained with the assumption of the data following a multivariate normal distribution.

A problem using validity constraints in the B2B optimisation is that it can often lead to infeasible problems in the MVT optimisation. This was found by looking at the poor performance and failed optimisations observed in a second case study shown in Appendix B and previous studies (Duran-Villalobos et al., 2016) when using validity constraints. A possible explanation for this issue is that the problem is overly constrained due to the changing conditions of the score space and the variability in the raw materials from one batch to the next.

6.2. Missing data algorithms in the B2B optimisation campaign

Another interest of the work presented in this article is to compare the use of different missing data algorithms in the estimation of the end-point quality over the B2B optimisation campaign.

Fig. 7 shows the average final penicillin concentration of 30 replicates, for 2 different missing data algorithms (PMP and TSR),

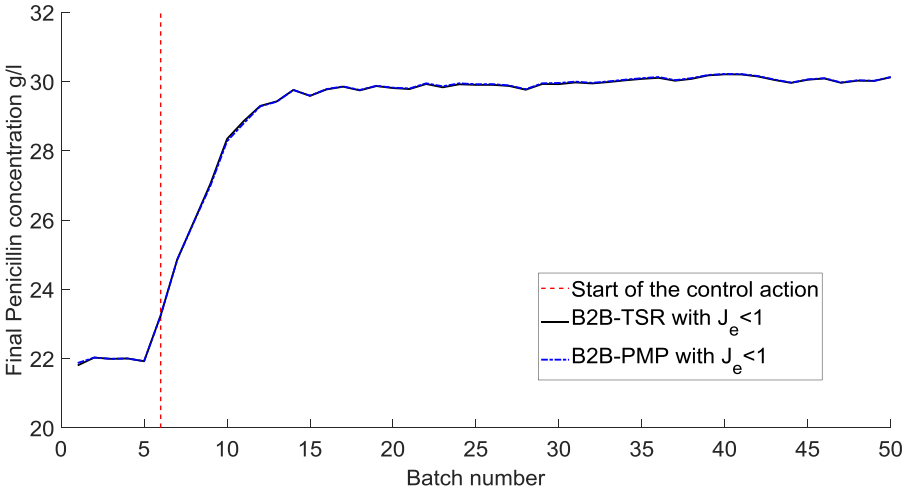

Fig. 7. Final penicillin concentration mean for the B2B campaign under different missing data algorithms.

Table 5
Evaluation parameters for the B2B campaign under different missing data algorithms.

Control methodology	Yield: mean of batch 50	Dispersion: standard deviation of batch 50	Batch at which convergence achieved
B2B-PMP	30.13 g/l	1.64 g/l	≈Batch 15
B2B-TSR	30.12 g/l	1.63 g/l	≈Batch 15

of an IndPenSim B2B optimisation campaign. Both methodologies use the validity constraint $J_e < 1$ without the assumption of multivariate normal distribution in the dataset. Additionally, Table 5 shows a series of metrics associated with the results displayed in Fig. 7.

Fig. 7 and Table 5 show no significant difference in the results obtained using PMP or TSR when applied to the QP optimiser in the B2B campaign. This similarity in these results could be attributed to the equivalence of both methods when the residuals are negligible.

The results for the B2B campaign show a significant improvement in the yield. This improvement goes from a mean value of 21.84 g/l to approximately 30 g/l. The improvement also occurs very quickly, with the yield increasing sharply for the first 10 batches after the B2B campaign starts (5 batches after the control action starts), and then a gradual improvement of the yield over several more batches. Similarly, in Fig. 3 the volume reaches the maximum volume allowed by the optimisation by batch 10. This suggests a strong link between the volume in the MVT and the yield, which is to be expected.

6.3. Missing data algorithms in the MPC campaign

As previously stated, the MPC campaign had the objective to reduce the batch-to-batch variation in the final penicillin concentration under the presence of initial variability in the raw materials and in-batch fluctuations in the process. This objective was achieved by repeatedly taking measurements and optimising the trajectory of the sugar feed flow rate during the batch. For example, Fig. 8 shows the progression of the substrate feed rate during a typical batch of the MPC campaign. The graph shows only small changes were made to the 'golden trajectory' during this batch. The golden trajectory was the optimal feeding profile suggested in (Goldrick et al., 2015).

The final penicillin concentration measurements made during 80 batches when MPC was applied, were collected. In each batch, 18 control points were applied. The control action started 50 h after the start of each batch and was repeated every 10 h until the end of the batch. The QP optimisation used at each control point used validity constraints on both T^2 and Q, as the detrimental effect of the T^2-statistic-based validity constraint, present in the B2B campaign, was not observed in the MPC campaign. The validity constraints were defined without assuming that there was a multivariate normal distribution in the dataset.

Table 6
Evaluation parameters for the B2B campaign under different missing data algorithms.

Control methodology	Yield average	MSE from the set-point
No control	29.48 g/l	1.32 g/l
MPC-PMP	30.05 g/l	0.99 g/l
MPC-TSR	29.99 g/l	1.08 g/l

Fig. 9 shows the final penicillin concentration of an IndPenSim MPC campaign when both of the missing data algorithms (PMP and TSR) were applied for a typical batch. This graph show a gradual drop in the variability of the final penicillin concentration along the MPC campaign.

Table 6 highlights several metrics taken from the results displayed in Fig. 9. The first column presents the final penicillin concentration during the MCP campaign. This metrics record any bias from set-point that might exist during the MPC campaign. The second column presents the Mean Square Error (MSE) of the actual end-point quality relative to the set-point during the MPC campaign. This parameter is a measure of the dispersion from the set-point during the MPC campaign.

The results in Table 6, along with the results from Fig. 9, reveal that there is no significant difference using the PMP or the TSR algorithms in the QP optimisation. The results also show a clear improvement in the average yield and the MSE in the campaigns where MPC was applied.

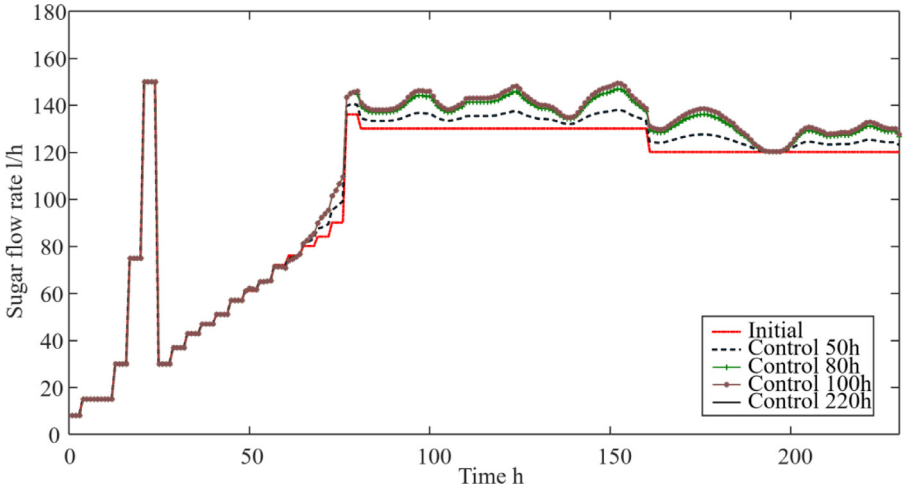

Fig. 8. 'Golden' MVT progression for a typical batch using MPC.

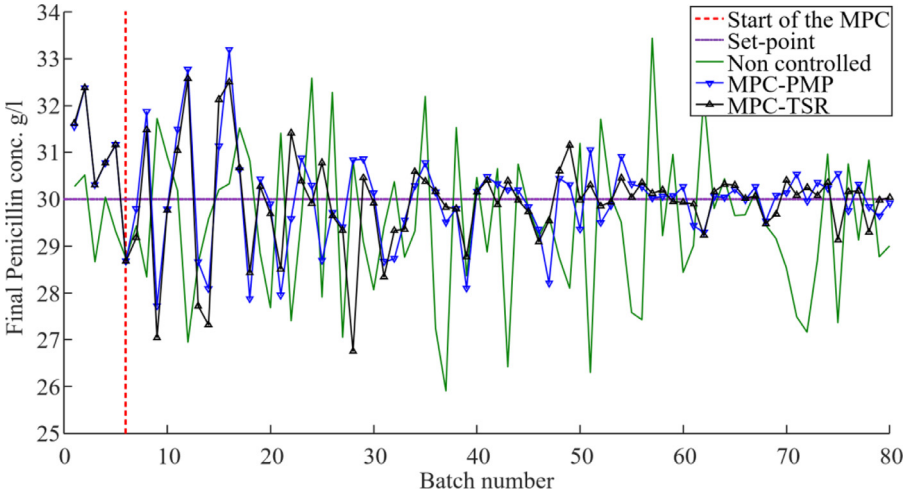

Fig. 9. Final penicillin concentration for the MPC campaign under different missing data algorithms.

6.4. MPC campaign after B2B campaign

The results shown in this section provide a comparison of the performance of the MPC campaign using the 'golden trajectory' identified in Goldrick et al. (2015) with the final trajectory that was determined following the B2B-TSR campaign. The objective of this comparison is to observe the effect of using an MPLS-model with an extensive set of batches and the reproducibility of the MPC performance when using different 'golden' trajectories.

Fig. 10 compares the final penicillin concentration of an IndPenSim MPC campaign using the 'golden trajectory' feed from Section 6.3 against the feed trajectory that was optimised using the B2B-TSR approach. To compare approaches using the same seeds for the random number generator, the batch number shown in Fig. 10 starts from batch 6 since MPC-TSR requires 5 batches of data to initialise the model. The graph shows there is slightly less variability at the beginning of the MPC campaign, after the process was optimised using the B2B-TSR technique.

Table 7 show the same evaluation parameters from Table 6 taken from the results displayed in Fig. 10. This table shows no significant difference in the yield from the two approaches. However, it highlights the reduction in MSE that results following B2B-TSR optimisation.

Table 7
Evaluation parameters for the B2B campaign under different missing data algorithms.

Control methodology	Yield average	MSE to the set-point
MPC-TSR	29.99 g/l	1.08 g/l
MPC-TSR + B2B-TSR	29.92 g/l	0.69 g/l

What stands out from the MPC campaign results is a slight improvement in the final penicillin concentration mean, by moving from values of 29.48 g/l in nominal runs to values very close to 30 g/l. Similarly, the consistency when applying MPC was substantially improved by reducing the MSE to the set-point from values of 1.32 g/l in nominal runs to values close to 1 g/l when the MPC was applied.

The MSE was further improved to 0.69 g/l when using the dataset from a B2B campaign beforehand. A likely explanation for this is that the MPLS model is much more accurate for the first MPC runs. This can be inferred from Fig. 10, where the MPC-TSR tech-

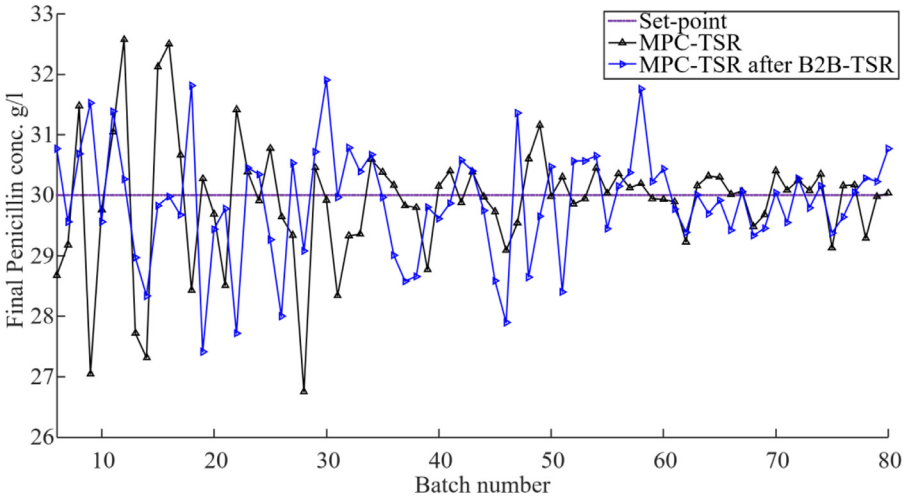

Fig. 10. Final penicillin concentration for the MPC campaign under different control strategies.

nique, using the B2B-TSR campaign dataset, shows much less variability than the one without the B2B-TSR campaign, whereas the variability at the end of the MPC-campaign shows little difference with both approaches.

7. Conclusions

In this article, an improved B2B optimisation strategy was successfully implemented on an industrial fed-batch penicillin simulation, greatly improving the yield from a nominal pre-optimised feed trajectory. The results showed that this control strategy converges to an optimal MVT, reaching the desired end-point quality consistently, after 10 batches, with only 5 batches used to identify the initial model.

An innovative Model Predictive Control strategy was successfully applied to the same simulation. This controller brought the values of the yield closer to the set-point and reduced process variability along multiple runs. The main advantage of this control strategy was its ability to reduce the influence of B2B variation in the quality of the raw materials and process variation through adjustments in the feeding strategy rate at multiple times along a batch.

Regarding the performance of the different missing data algorithms that were applied when optimising the MVT using QP, the results did not show a considerable difference when using TSR or PMP to estimate the end-point quality through the batch. However, PMP has a more straightforward interpretation and requires less computing power.

With respect to the benefits of applying the proposed confidence limits in the validity constraints of the QP, the results were improved in the B2B campaign when using a bootstrap calculation than when using other literature approaches which considered the dataset to have a multivariate normal distribution. This finding suggests that applying the bootstrap calculation in the validity constraints offers a more robust approach than other techniques which typically require tuning. In spite of these results, previous findings and a second case study shown in Appendix B suggest that the use of validity constraints in the B2B optimisation and varying initial conditions in the raw materials can lead to infeasible QP problems.

This study suggests that the application of the proposed control strategies, together or individually, to an industrial fed-batch process would lead to improved consistency and yield in the existence of plant and raw materials variability.

Funding

This work was supported by the UK Engineering & Physical Sciences Research Council (EPSRC) [EP/P006485/1] and a consortium of industrial users and sector organizations in the Future Targeted Healthcare Manufacturing Hub hosted by UCL Biochemical Engineering in collaboration with UK universities.

Declaration of Competing Interest

None.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2019.106620.

References

Arteaga, F., Ferrer, A., 2002. Dealing with missing data in MSPC: several methods, different interpretations, some examples. J. Chemom. 408–418. doi:10.1002/cem.750.

Birol, G., Undey, C., Cinar, A., 2002. A modular simulation package for fed-batch fermentation: penicillin production. Computers and Chemical Engineering. 26, 1553–1565. doi:10.1016/S0098-1354(02)00127-8.

Bonvin, D., 1998. Optimal operation of batch reactors - a personal view. J. Process Control 8, 355–368.

Bonvin, D., Srinivasan, B., Hunkeler, D., 2006. Control and optimization of batch processes: improvement of process operation in the production of specialty chemicals. IEEE Control Syst. Mag. 34–45. doi:10.1109/MCS.2006.252831.

Camacho, J., Lauri, D., Lennox, B., Escabias, M., Valderrama, M., 2015. Evaluation of smoothing techniques in the run to run optimization of fed-batch processes with u-PLS. J. Chemom. 29, 338–348. doi:10.1002/cem.2711.

Camacho, J., Pico, J., Ferrer, A., 2007. Self-tuning run to run optimization of fed batch processes using unfold PLS. AIChE J. 53. doi:10.1002/aic.

Dayal, B.S., MacGregor, J.F., 1997. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. J. Process Control 7, 169–179. doi:10.1016/S0959-1524(97)80001-7.

Desharnais, B., Camirand-Lemyre, F., Mireault, P., Skinner, C.D., 2015. Determination of confidence intervals in non-normal data: application of the bootstrap to cocaine concentration in femoral blood. J. Anal. Toxicol. 39, 113–117. doi:10.1093/jat/bku127.

Duran-Villalobos, C.A., Lennox, B., Lauri, D., 2016. Multivariate batch to batch optimisation of fermentation processes incorporating validity constraints. J. Process Control 46, 24–42. doi:10.1016/j.jprocont.2016.07.002.

Ed, P.P., 2013. LNAI 7987- Advances in data mining. 13th Industrial Conference, ICDM doi:10.1007/978-3-642-39736-3.

Flores-Cerrillo, J., MacGregor, J.F., 2003. Within-batch and batch-to-batch inferential-adaptive control of semibatch reactors: a partial least squares approach. Ind. Eng. Chem. Res. 42, 3334–3345. doi:10.1021/ie020596u.

Flores-Cerrillo, J., MacGregor, J.F., 2004. Control of batch product quality by trajectory manipulation using latent variable models. J. Process Control 14, 539–553. doi:10.1016/j.jprocont.2003.09.008.

Flores-Cerrillo, J., MacGregor, J.F., 2005. Latent variable MPC for trajectory tracking in batch processes. J. Process Control 15, 651–663. doi:10.1016/j.jprocont.2005.01.004.

García-Muñoz, S., Kourti, T., MacGregor, J.F., 2004. Model predictive monitoring for batch processes. Ind. Eng. Chem. Res. 43, 5929–5941. doi:10.1021/ie034020w.

Gins, G., Vanlaer, J., Van Impe, J.F.M., 2009. Online batch-end quality estimation: does laziness pay off? IFAC Proc. Vol. doi:10.3182/20090630-4-ES-2003.0321.

Goldrick, S., Andrei, S., Lovett, D., Montague, G., Lennox, B., 2015. The development of an industrial-scale fed-batch fermentation simulation. J. Biotechnol. 193, 70–82. doi:10.1016/j.jbiotec.2014.10.029.

Jackson, J.E., Mudholkar, G.S., Edward, J., 1979. Control procedures for residuals associated with principal component analysis. Technometrics 21, 341–349. doi:10.2307/1267757.

Joe Qin, S., 1998. Recursive PLS algorithms for adaptive data modeling. Comput. Chem. Eng. 22, 503–514. doi:10.1016/S0098-1354(97)00262-7.

Kumar, A.S., Ahmad, Z., 2012. Model predictive control (MPC) and its current issues in chemical engineering. Chem. Eng. Commun. 199, 472–511. doi:10.1080/00986445.2011.592446.

Laurí, D., Lennox, B., Camacho, J., 2014. Model predictive control for batch processes: ensuring validity of predictions. J. Process Control 24, 239–249. doi:10.1016/j.jprocont.2013.11.005.

Laurí, D., Sanchis, J., Martínez, M., Hilario, a, 2013. Latent variable based model predictive control: ensuring validity of predictions. J. Process Control 23, 12–22. doi:10.1016/j.jprocont.2012.11.001.

Liu, K., Chen, Y., Zhang, T., Tian, S., Zhang, X., 2018. A survey of run-to-run control for batch processes. ISA Trans. doi:10.1016/j.isatra.2018.09.005.

Luo, L., Bao, S., 2018. Knowledge-data-integrated sparse modeling for batch process monitoring. Chem. Eng. Sci. 189, 221–232. doi:10.1016/j.ces.2018.05.055.

Martens, H., 2001. Reliable and relevant modelling of real world data: a personal account of the development of PLS regression. Chemom. Intell. Lab. Syst. 58, 85–95. doi:10.1016/S0169-7439(01)00153-8.

Martens, H., Naes, T., 1989. Multivariate Calibration. Wiley.

Nelson, P., Taylor, P., 1996. Missing data methods in PCA and PLS: score calculations with incomplete observations. Chemom. Intell. Lab. Syst. 35, 45–65. doi:10.1016/S0169-7439(96)00007-X.

Nomikos, P., Macgregor, J.F., 1995a. Multivariate SPC charts for batch monitoring processes. Technometrics 37, 41–59. doi:10.1080/00401706.1995.10485888.

Nomikos, P., MacGregor, J.F., 1995b. Multi-way partial least squares in monitoring batch processes. Chemom. Intell. Lab. Syst. 30, 97–108. doi:10.1016/0169-7439(95)00043-7.

Qin, S.J., 2003. Statistical process monitoring: basics and beyond. J. Chemom. 17, 480–502. doi:10.1002/cem.800.

Qin, S.J., 2012. Survey on data-driven industrial process monitoring and diagnosis. Annu. Rev. Control 36, 220–234. doi:10.1016/j.arcontrol.2012.09.004.

Ramani, S., 1974. EDF statistics for goodness of fit and some comparisons. J. Am. Stat. Assoc. 69, 730–737. doi:10.2307/2286009.

Sheng, J., Chen, T., Shah, S.L., 2002. GPC for non-uniformly sampled systems based on the lifted models. IFAC Proc. Vol. 405–410. doi:10.1016/S0959-1524(02)00009-4.

Ündey, C., Ertunç, S., Çinar, A., 2003. Online batch/fed-batch process performance monitoring, quality prediction, and variable-contribution analysis for diagnosis. Ind. Eng. Chem. Res. 42, 4645–4658. doi:10.1021/ie0208218.

Vanlaer, J., Van Den Kerkhof, P., Gins, G., Van Impe, J.F.M., 2011. The inuence of measurement noise on PLS-based batch-end quality prediction. IFAC Proc. Vol. 7156–7161. doi:10.3182/20110828-6-IT-1002.02775.

Wan, J., Marjanovic, O., Lennox, B., 2012. Disturbance rejection for the control of batch end-product quality using latent variable models. J. Process Control 22, 643–652. doi:10.1016/j.jprocont.2011.12.012.

Wold, S., Geladi, P., Esbensen, K., Öhman, J., 1987. Multi-way principal components-and PLS-analysis. J. Chemom. 1, 41–56. doi:10.1002/cem.1180010107.

Xu, Q.-S., Liang, Y.-Z., 2001. Monte Carlo cross validation. Chemom. Intell. Lab. Syst. 56, 1–11. doi:10.1016/S0169-7439(00)00122-2.

Yabuki, Y., Nagasawa, T., Macgregor, J.F., 2000. An industrial experience with product quality control in semi-batch processes. Comput. Chem. Eng. 24, 585–590. doi:10.1016/S0098-1354(00)00423-3.

Yu, L.X., Amidon, G., Khan, M.A., Hoag, S.W., Polli, J., Raju, G.K., Woodcock, J., 2014. Understanding pharmaceutical quality by design. AAPS J. 16, 771–783. doi:10.1208/s12248-014-9598-3.